Author:
Tatton-Brown, Oliver M W

Title:
Rigour, Proof and Soundness

Rigour, Proof and Soundness

By
Oliver Tatton-Brown

Department of Philosophy
School of Arts
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Arts.

March 2020

Word count: 79746

Abstract

The initial motivating question for this thesis is what the standard of rigour in modern mathematics amounts to: what makes a proof rigorous, or fail to be rigorous? How is this judged? A new account of rigour is put forward, aiming to go some way to answering these questions. Some benefits of the norm of rigour on this account are discussed. The account is contrasted with other remarks that have been made about mathematical proof and its workings, and is tested and illustrated by considering a case study discussed in the literature.

On the view put forward here one can obtain a manner of informal, rigorous mathematics founded on any of a variety of proof systems. The latter part of the thesis is concerned with the question of how we should decide which of these competing proof systems we should base our mathematics on: i.e., the question of which proof system we should take as a foundation for our mathematics. A novel answer to this question is proposed, in which the key property we should require of a proof system is that for as many different kinds of structures as possible, when the proof system allows a generalization about that kind of structure to be proved, the generalization actually holds of all real examples of that kind of structure which exist. This is the requirement of soundness of the proof system (for each kind of structure). It is argued that the best way to establish the soundness of a proof system may be by giving an interpretation of its axioms on which they are established as true. As preparation for this discussion, the thesis first investigates the logical and conceptual basis of various arithmetic concepts, with the results obtained used in the final discussion of soundness.

Dedication and Acknowledgements

To Alex and Amy

I would like to extend my sincere gratitude to a number of people who have made this work possible. Leon Horsten has been a wonderful guide into academic philosophy, the source of too many fascinating conversations to count, and has been admirably patient and nurturing with my various proto-ideas. Catrin Campbell-Moore has done much sterling work identifying places where the thesis was lacking in clarity or force. In a few key conversations Philip Welch has made various remarks that have contributed greatly. As well as the above, a variety of people have provided very helpful (and often very detailed) comments on my papers, including in particular Daniel Waxman, Brendan Larvor, Marcus Giaquinto, Yacin Hamami, and various anonymous referees. I am also thankful for crucial remarks from members of the audience at talks I have given, including at the FSB research seminars in Bristol, and at conferences in Sicily, London Ontario, Cambridge, and Leuven. Finally I would like to thank Mungo for being my initial companion in my first days of interest in philosophy, and Alex and Amy for their endless help and support. This work was generously funded by the AHRC, grant AH/L503939/1, via the SWWDTP.

Chapter II is an edited version of Tatton-Brown (2019b), and chapter IV is an edited version of Tatton-Brown (2019a).

Author's Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .....

DATE: .....

Contents

List of Figures

Introduction

A mathematical proof of a proposition is often taken to be amongst the best kinds of evidence for it, but is this appropriate? What can be said about the standard of proof in mathematics – what is it that makes a given argument valid or not? Can we give reasons why this is a good standard of evidence? The first part of this thesis addresses these questions, at least for rigorous proof – the standard of valid proof applied in much of modern mathematics. Chapter I gives a new account of mathematical rigour, building on the work of Burgess (2015). Some benefits of rigour on this account are discussed: in particular, all rigorous arguments are formalizable in principle, and one has a reliable mechanism for resolving disputes between mathematicians about the validity of arguments.

Many authors object to the kind of account of rigour put forward in chapter I, in which there is a link between rigour and formalizability. Some objections are considered in chapter I, but a remaining concern that many have expressed is that informal proofs may consist of a kind of irreducibly high level reasoning, strongly resistant to formalization, requiring radically new ideas perhaps to be introduced before it can be formalized. Though a number of authors have expressed a concern along these lines, rarely do they put forward substantial examples of arguments that exhibit the property they claim. An exception is an argument from knot theory known as Alexander's lemma (Alexander 1923), which was recounted by Field's medallist Vaughan Jones in a philosophical piece (Jones 1998) in which he claims that it is an easy result that would be very difficult to

formalize. De Toffoli and Giardino (2016) use the argument as the basis of a case study in which they defend a rather different view of proof to that in chapter I, with a number of their claims reiterated by Larvor (2019).

Clearing up these various claims about Alexander's lemma is the primary purpose of chapters II and III. There are now multiple different versions of the argument in play: Alexander's original, which works with polygonal knots, Jones's retelling which works with smooth knots, and De Toffoli and Giardino's version, which combines aspects of both. Each merits a different response. Chapter II argues that despite De Toffoli and Giardino's claims, Alexander's original argument is perfectly rigorous by the normal standard, as enunciated in chapter I; all the properties they attribute to it which clash with this standard are actually only found in their retelling of the argument, rather than the original, and are inherited from Jones's version of the argument, a consequence of them combining these two versions of the arguments together. Noting these points removes the basis for the view of proof that De Toffoli and Giardino advocate, and nullifies their objections to the kind of view of proof seen in chapter I. The remaining issue is the status of Jones's version of the argument. Chapter III argues that though the argument as he tells it would indeed be very hard to formalize, this is simply because it is deeply flawed from a rigorous point of view, as rigour is normally understood. The argument for this illustrates the closeness of the view of rigour given in chapter I to standard features of the concept. It is also an opportunity to address the question of how pictorial arguments fit into the picture given in chapter I; many authors feel that pictorial arguments may give good examples of valid arguments that resist formalization, and Jones and De Toffoli and Giardino both emphasise the visual features of their versions of Alexander's lemma. Chapter III uses Jones's version of the argument to argue for a general characterization of what rigour requires of pictorial arguments, expanding on the work of Larvor (ibid.), and argues that this defuses the threat that pictorial arguments had been thought to pose to a link between rigour and formalizability.

The account of rigour in chapter I uses the current practice of mathematics as based on ZFC as its focus, but one could found one's rigorous mathematics on any of a variety of proof systems, leading to different options for what reasoning would be available. Thus the question of which proof system we should use as a basis arises: this is the question of which foundation we should choose for our mathematics. The latter part of the thesis is ultimately concerned with this question, with chapter VI arguing that when considering whether a given proof system could be used as a foundation for mathematics, the key property we should require for it is that it be *sound* for as many different kinds of mathematical structures as possible: i.e., that for as many different kinds of mathematical structures as possible, when a generalization about that kind of structure is derivable in the proof system, that generalization actually holds of all real examples of that kind of structure which exist.

It is argued that many obvious routes to try to establish the soundness of a proof system will not work, and that in fact, the best route we have to establishing the soundness of a general proof system is via the iterative conception of set, or the limitation of size conception: if one of these accounts – or a combination, or some other account – justifies the axioms of set theory, then we can argue that set theory is sound, and thus a potential foundation. Thus we have a case for *veritism*, being defined here (in the context of philosophy of mathematics) as the view that it matters whether the principles we use in mathematics have some subject matter they are true of. ETCS and homotopy type theory may then have their soundness at least partially justified by giving an interpretation of them in set theory, and proving a relative soundness result. These findings are used to argue against the view of set theory and extrinsic evidence put forward by Maddy (2011).

Key to the discussion of soundness is the ability to give characterizations of mathematical structures that have real world content, that are the kind of thing that can be satisfied or fail to be satisfied by real world objects (as opposed to being purely formal

definitions); and to state real properties that such structures may have. Our ability to do these things is dependent on what logical resources we take there to be available, logic being the way we model our language's ability to state definitions and properties. As preparation for the discussion in chapter VI, the two short chapters chapters IV and V discuss the conceptual and logical resources underlying various arithmetic concepts. Chapter IV considers what is required to define functions by primitive recursion, arguing that a logical operator called the double ancestral captures what is conceptually required by this ability in a satisfying way. As an incidental application, this is used to strengthen an argument of Smith (2008) for Isaacson's thesis. Then chapter V argues that combining the ancestral and double ancestral operators with plural logic allows a natural way to define the concepts of finiteness and equinumerosity for finite pluralities, as well as the operations of addition and multiplication for finite pluralities. These are used to sketch a new interpretation of arithmetic, and aspects of this are contrasted with the Neo-Fregean interpretation.

Chapter I

Understanding Rigour

1 Questions about proof

What's going on in mathematical proofs? How do they establish the truth of their conclusions? By *proof* I mean the kind of proof mathematicians actually write and exchange with each other and accept as valid.

One way to attempt to understand proof is via derivations, the formal objects that logicians use to model deduction. Gentzen intended his system of natural deduction to be

> a formalism that reflects as accurately as possible the actual logical reasoning involved in mathematical proofs (Gentzen 1969, p. 74)

and he described its derivations as having

> [a] close affinity to actual reasoning (ibid., p. 80).

One can read this as meaning that all valid mathematical inferences should be instances of the logical rules of natural deduction, or closely related to them.[1] If mathematicians

[1] This is apparently the reading of Goethe and Friend (2010, pp. 274–275). It is also possible that Gentzen merely intended that the logical substructure of proofs should be expressible in natural deduction – aspects such as introducing and eliminating premises, moving from a statement about an arbitrary x to a statement about all x, and so on.

did actually explicitly work according to certain fixed formal rules, then proof would be unmysterious: understanding proof would just come down to understanding the relevant formal rules. However as many authors point out, in reality mathematical proofs do not proceed according to any list of rules that one could specify in advance.[2] There is far too much variety of inferences for that, and new proofs will often contain inferences that are somewhat (or completely) novel.

More plausible than a naive rules based view of proof is one on which the correctness of a proof consists in its being *formalizable* – translatable into a derivation according to some given system of formal rules. What exactly this description amounts to will depend on what notion of translation is employed (and on the underlying system of formal rules). Though historically often implicitly assumed by philosophers of mathematics to be correct, numerous authors have recently objected to this view. There have thus been many calls to develop a new, more plausible account of mathematical proof and its epistemology.[3]

This chapter aims to provide such an account, or at least an outline of it – found largely in sections I.3 and I.4. The account is not actually of proof in general, but only of *rigorous* proof, rigour being the standard of acceptable proof in much of modern mathematics. I do not view this as a significant limitation. Indeed I argue in section I.2 that questions like "what is mathematical proof?", asked in full generality, are unlikely to receive a satisfying answer: there is no univocal notion of proof in mathematics, or at least not one we can expect to obtain a substantial philosophical analysis of. The account of rigorous proof given here is not the last word on the subject, and I note places where it could be expanded on. In chapter II and chapter III case studies are used to test the account and contrast it with alternatives.

[2]See for instance Tragesser (1992), Celluci (2009), Leitgeb (2009), Goethe and Friend (2010), and Larvor (2012).

[3]As illustrations of the objections, and the calls for improvement, see for instance Rav (1999; 2007), Celluci (2009), Detlefsen (2009), Leitgeb (2009), Pelc (2009), Goethe and Friend (2010), Antonutti Marfori (2010), Larvor (2012), Weir (2016), De Toffoli and Giardino (2016), and Larvor (2019).

The account put forward here takes a somewhat novel approach: instead of focusing exclusively on proofs, considering also the ability of mathematicians to produce and recognize valid inferences. Where does this ability to recognize validity come from? What can be said about it? Answering these questions is one way to gain insight into what is going on in proofs themselves - how they justify their conclusions.

In this chapter the resulting account is used to address two questions about proof that have been raised in the literature. The first is that of how mathematicians are so able to generate agreement about the validity of proofs. This phenomenon – that if a mathematician thinks a proof is valid, they can generally convince others, or be convinced themselves of a flaw in it – has been noted by various authors, including Azzouni (2004, pp. 83–84) and Antonutti Marfori (2010, pp. 267, 270–271). Explaining it is one of the major motivations for the "derivation–indicator" view of proof that Azzouni (2004) puts forward. Azzouni's analysis of proof has met with controversy, with for instance Tanswell (2015) raising what appear to be valid criticisms. Using the account of rigour put forward here, in section I.5 I give an explanation of this agreement about validity that aims to be simpler and more plausible than Azzouni's.

The second question is that of whether rigorous proofs are necessarily formalizable. Many authors answer this question in the negative, for a variety of different reasons, some of which are considered here. One cannot properly discuss formalizability without discussing what formal system one is targeting, and the first objection to formalizability considered is the claim that mathematicians should really be regarded as working in naive set theory, with its unrestricted comprehension scheme. This has been claimed for instance by Jones (1998) and Leitgeb (2009). If correct this would render the formalizability claim empty of interest, since set theory with unrestricted comprehension is inconsistent and so any argument is trivially formalizable in it (with every inference justified by your favourite set theoretic paradox). This position is evaluated and dismissed in section I.6.

Next, the family of worries stemming from Rav (1999) is considered. He argues that when filling in intermediate steps in an argument, we have no reason to believe the process will necessarily terminate. This worry is reiterated by Weir (2016), who uses it as motivation for a version of formalism which can handle infinitary proofs. Pelc (2009) raises a weaker version of the worry, arguing that the process of filling in intermediate steps will terminate in principle, but that we have no reason to believe it will terminate in a proof of feasible length, say involving less than 10^{120} steps. In section I.7 a response to Rav's worry is put forward, and a possible response to Pelc's worry also sketched.

A third kind of worry about formalizability asks how, if part of our norm for validity is that arguments they be formalizable, are mathematicians so good at judging this? After all, mathematicians generally seem to be reliable at judging validity, with arguments accepted as valid generally staying that way, and agreement about validity often quickly reached (as noted above). If part of our norm for validity is that proofs be formalizable, there is a mystery about how mathematicians are so good at judging this – since most mathematicians do not know the axioms of set theory, are unfamiliar with the rules of logic, have never used a proof assistant, and in general have no real experience of or interest in the process of formalization. This is also addressed in section I.7.

These are not all the worries about formalizability that have been raised. Another prominent family of worries maintains that a proof may involve irreducibly high level reasoning – reasoning that is very difficult to formalize, that may require radically new ideas for its formalization, that may not be "the same argument" when formalized (Rav 1999; 2007; Celluci 2009; Leitgeb 2009; Goethe and Friend 2010; Larvor 2012). Chapters II and III consider in detail one strain of this kind of thought.

2 Initial remarks on rigour

The main account of rigour is found in sections I.3 and I.4 and, but first there are some preliminary remarks worth making. To begin with, we will see some examples of non rigorous mathematics; this helps illustrate the distinctive nature of mathematical rigour, and is also used to argue against the idea that there is a unified notion of "proof" in mathematics that is worth conceptually analyzing. Indeed one of the intended lessons of this chapter is that it is rigorous proof, not proof in general, which is the philosophically interesting concept. The latter part of this section summarizes the discussion of rigour given by Burgess (2015), which makes a number of valid, significant points, but which does not address the questions discussed in section I.1 that this chapter attempts to answer.

First, the examples of non rigorous mathematics. One good example consists of manipulations involving infinitesimals in the 17^{th} and 18^{th} centuries, which – before the introduction of limits into analysis – were not generally rigorous. For a toy example of how they often worked, we can determine the derivative of the function $x \mapsto x^2$. If we let dx be small, then we have

$$\frac{(x+dx)^2 - x^2}{dx} = \frac{x^2 + 2xdx + dx^2 - x^2}{dx} = \frac{2xdx + dx^2}{dx} = 2x + dx$$

and then since dx is small we discard it, obtaining $2x$ as the derivative of $x \mapsto x^2$ at x. Arguments along these lines (and more complicated versions) were carried out by various authors, with Fermat perhaps being the first to give this particular kind of calculation (Kline 1990b, pp. 344–345). These methods met with criticism however, as it was not clear what the status of "small" quantities such as dx was, or what was allowed when manipulating them. Indeed if dx is small but non zero then the result is only approximate; for the result to be exact, we require that $2x + dx = 2x$, but then we obtain by subtraction that $dx = 0$ and so we cannot divide by dx to begin with.

Rolle pointed this out (Mancosu 1989, pp. 230–231), followed more famously by Berkeley who complained that infinitesimals appeared to be the "ghosts of departed quantities" (Berkeley 1999, pp. 80–81).

Other common methods in the 17$^{\text{th}}$ and 18$^{\text{th}}$ centuries also lacked rigour. Often arguments proceeded by assuming that what held for the finite also held for the infinite, with infinite series being manipulated as though they were finite sums, without worrying about issues of convergence. For instance Jacob Bernoulli argued (essentially) that

$$\begin{aligned} 1 &= \left(1 + \frac{1}{2} + \frac{1}{3} + \dots\right) - \left(\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots\right) \\ &= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \dots = \frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \dots \end{aligned}$$

despite knowing that the sum $1 + \frac{1}{2} + \frac{1}{3} + \dots$ is infinite (Kline 1990b, p. 443).[4] As illustrated by many further examples in Kline (ibid., Chapter 20), infinite series were freely manipulated in this period without worrying about convergence, despite these methods sometimes leading to false or contradictory conclusions.

Arguments like these gradually came to be seen as unacceptable. In the 19$^{\text{th}}$ century both the calculus and the study of infinite series were rephrased in terms of the concept of limit, putting them on a firm footing (Kline 1990a, Chapter 40). Infinitesimals were then largely eschewed in analysis until Robinson demonstrated how one could in fact reason rigorously about them, via the logical concept of a non standard model (Robinson 1996). Recent work has shown how the inconsistency in arguments like the above involving infinitesimals can be embraced and utilized, in a suitable paraconsistent logic (see for instance Brown and Priest 2004, and Sweeney 2014).

Other ways of reasoning formerly regarded as valid also came to be shunned, such

[4] One can adjust this argument to make it rigorous by telescoping the partial sums:

$$\frac{1}{2} + \frac{1}{6} + \dots + \frac{1}{n(n+1)} = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \dots + \left(\frac{1}{n} - \frac{1}{n+1}\right) = 1 - \frac{1}{n+1}$$

as appeals to intuition. Proofs were demanded of even very intuitive statements, with the Jordan curve theorem being a famous example. This states that if $\phi : S^1 \to \mathbb{R}^2$ is continuous and injective then $\mathbb{R} \setminus \phi(S^1)$ consists of exactly two connected components, one of which is bounded and the other unbounded, and $\phi(S^1)$ is the boundary of each component. If one explained all the relevant terms – "continuous", "injective", "bounded" and so on – to a layman, they would likely state that this was simply immediate, and one might well get the same reaction from some undergraduates (particularly those with a preference for applied mathematics). It takes some effort to imagine a curve ϕ in such a way that the conclusion seems anything but obvious. Nonetheless mathematicians were not satisfied with this, and providing a rigorous proof turned out to be very difficult. Bolzano had already noted that this fact required proof (Coulston 1970, p.274) and this proof was only provided by Jordan (published in Jordan 1887) more than 30 years after Bolzano's death.[5]

There are a number of lessons to be drawn from these examples. Firstly, they can be used to assess the adequacy of an account of rigour, such as that given in sections I.3 and I.4 – it needs to be able to explain why they are not considered rigorous. Secondly, such examples illustrate that an argument does not have to be rigorous to be reliable, or explanatory, or valuable. Arguments involving infinitesimals in the 18^{th} century could plausibly be all three (such as the differentiation example above), as could manipulations of infinite series. Much of modern non rigorous physics and engineering is also presumably reliable, explanatory and valuable. Nonetheless rigour does bring benefits, some of which will be discussed in this chapter.

These examples also illustrate the broadness of the notion of mathematical proof in times past. This, I believe, tells against the desire to seek a philosophical account of the general notion of proof – to discover what proof in general "really is", or where

[5]Some controversy followed Jordan's proof, with Veblen (1905) claiming it was flawed, and claiming to give the first rigorous proof. However Hales (2007) argues that Jordan's proof was basically valid, though perhaps not as polished as it could be.

its boundaries are drawn. Indeed in the above examples, reasoning of various kinds all distinct from the usual mathematical paradigm of deduction is seen. To make arguments involving infinitesimals one postulates a new manner of calculation, in which a quantity is treated at one stage as non zero, and later as small enough that it can be neglected. This can otherwise be viewed as the postulation of a new kind of entity with these apparently odd properties. Either way, it is essentially a form of abductive reasoning: reasoning of a kind which is not justified by anything that has gone before, but instead by its immense success in solving all manner of differential problems. It is not so different to the postulation of new principles or entities in physics, except that it is confirmed by mathematical applications and deductions, rather than by experiments. Then the manipulation of infinite series as though they were finite is essentially a case of argument by analogy, again backed up by its apparent success in solving problems. Finally we have appeals to intuition, delivering conclusions that one finds very hard to doubt because of one's intuitive grasp of the concepts involved – not so different from the intuition that philosophical zombies could exist, or that nothing can cause itself.

I do not think there is much to be gained by seeking to discover what these disparate forms of reasoning "have in common". They are all taken to justify high credence in their conclusions, and they all concern abstract, mathematical subject matter, but beyond that I am not sure there is much to be said. Certainly one could conduct a fruitful investigation into abductive reasoning, or intuition as a form of evidence, but there is unlikely to be much distinctive to say about either in the context of mathematical proof that does not apply to more general contexts. There is also not much I think to say about why such methods were accepted, beyond that they were felt to be reliable. The best assessment of proof in general may just be that there are different kinds of permissible actions that one may carry out, as Larvor (2012) puts it. It is *rigorous* proof which I think is more deserving of philosophical attention: this is where the ideal of flawless deduction that Euclid aspired to takes its purest form (Burgess 2015, pp. 36–38), and

where mathematical reasoning is found at its most distinctive, and epistemologically robust.

The variety of kinds of inference allowed in proofs historically makes it clear that it is only for rigorous mathematics that one could defend the formalizability of arguments. For instance what would a formal system for arguments like Bernoulli's, involving manipulation of divergent series, look like? Would it formalize analogies between the finite and the infinite? Similarly there is no reason to think that irreducibly intuitive reasoning would be formalizable. That it is only for rigorous mathematics that one could defend formalizability is a point that has perhaps not always been clearly grasped by objectors. Indeed in chapter III I analyze a knot theoretic proof sketch which has been claimed to be unformalizable (Jones 1998; De Toffoli and Giardino 2016), and argue that in fact the argument is far from rigorous, and this is the reason it fails to be formalizable.

Given the major differences between the historical standard of proof and what modern rigour permits, an obvious question is how and why the shift to rigour came about. Considering that would take us too far afield, but it is one topic which Burgess discusses (Burgess 2015, Chapter 1) in his account of rigour, to which we now turn. Burgess is mainly concerned with implications of the norm of rigour for the debate over structuralism, and does not give explicit arguments concerning the kinds of questions raised in section I.1 – the epistemology of proof, the ability of mathematicians to reach agreement about the validity of proofs, the issue of formalizability and so on. Nonetheless some observations he makes are worth highlighting.

Burgess emphasises that any piece of rigorous mathematics takes place in a context of existing results and definitions which can be appealed to (ibid., pp. 149–158). One can then extend the boundaries of knowledge with a new argument, stringing together definitions and proofs of propositions, appealing to existing results and using existing concepts where needed. Burgess also emphasises that often it does not matter whether a fact one is appealing to is a basic principle or a consequence thereof, or how concepts

used were actually defined as long as the properties one needs of them do hold. This is the basis of his critique of structuralism as a metaphysical position.[6]

Both aspects of the rigorous process – definitions, and proofs of propositions – merit some attention. Burgess notes that when introducing a new concept, one is required a clear definition in terms of existing ones (Burgess 2015, p. 7). This definition does not have to be completely formal – for instance one can state that a vector space is a set equipped with an abelian group structure and a scalar multiplication operation, without specifying exactly how this is coded set theoretically: as a triple $(V, +, \cdot)$, or $((V, +), \cdot)$, or as $(+, \cdot)$, or in some other way. It just needs to be clear that the definition could be made completely precise in such a way that all uses made of the concept would be valid.

Burgess also discusses what the standard of rigour requires for proofs of propositions. He considers various possibilities, and ultimately comes to the (tentative) conclusion that:

> What rigor requires is that each new result should be obtained from earlier results by presenting enough deductive steps to produce conviction that a full breakdown into obvious deductive steps would in principle be possible (ibid., p. 97)

This I think is basically right. However there is more to say before this can be brought to bear on the issues discussed in section I.1. If we are interested in the epistemology of proof, then this is only a sketch rather than a full account. How is this conviction generated? How is it reliable? If mathematicians are judging formalizability in principle (which is roughly what "full breakdown into obvious deductive steps" might amount to), how are they able to judge this? As discussed in section I.1, most mathematicians

[6]He argues that exactly how the concepts one uses were defined is often irrelevant, as all one will need are certain derived properties. Thus one need not care about how things were defined when doing mathematics. It is this irrelevance of definitions that Burgess argues has been mistaken by structuralists for a metaphysical truth about the nature of mathematical structures, with structuralists hoping to infer for instance that mathematical objects have only general structural properties (Burgess 2015, Chapter 3).

have no experience of or interest in formalization, after all. There is also the question of how mathematicians are so good at resolving disputes, discussed in section I.1, which we could hope to answer. Addressing these issues is the purpose of the remainder of the chapter.

3 A rigorous eduction

As mentioned in section I.1, the account of rigorous proof put forward here uses a somewhat novel approach: to try to understand the skill of mathematicians in judging and producing rigorous proofs by thinking about how this skill is acquired. For this we will start at the beginning. There are many different universities around the world that teach rigorous mathematics, and they may teach it in somewhat different ways, but there are some common features that can be pointed to. Students are generally taught the basics of rigorous proof by seeing and working through examples, paired with descriptions of how and why the reasoning involved works. An example of a basic early result students might see is displayed in fig. I.1.

2.1.8 Theorem **(a)** *If* $a \in \mathbb{R}$ *and* $a \neq 0$, *then* $a^2 > 0$.
(b) $1 > 0$.
(c) *If* $n \in \mathbb{N}$, *then* $n > 0$.

Proof. (a) By the Trichotomy Property, if $a \neq 0$, then either $a \in \mathbb{P}$ or $-a \in \mathbb{P}$. If $a \in \mathbb{P}$, then by 2.1.5(ii), $a^2 = a \cdot a \in \mathbb{P}$. Also, if $-a \in \mathbb{P}$, then $a^2 = (-a)(-a) \in \mathbb{P}$. We conclude that if $a \neq 0$, then $a^2 > 0$.
(b) Since $1 = 1^2$, it follows from (a) that $1 > 0$.

Figure I.1: Bartle and Sherbert (2000, p. 26) (proof of (c) is omitted)

This demonstrates a fact commonly assumed without question: that the square of a non zero real is positive. Probably the only part of the argument that requires explanation is the symbol $\mathbb{P}$, which denotes the set of (strictly) positive real numbers. Axioms concerning this set have been stated a few pages previously. The relevant axioms

are that:

(i) For any $a \in \mathbb{R}$, either $a \in \mathbb{P}$ or $(-a) \in \mathbb{P}$ or $a = 0$, with exactly one of these holding.

(ii) If $a, b \in \mathbb{P}$ then $a \cdot b \in \mathbb{P}$.

From these the above proof proceeds straightforwardly, arguing by cases.

The main thing to note about this proof is just how incredibly detailed it is. Virtually all of the logical structure of the argument is right there on the page. There are places where one could be even more explicit, in particular in the assertion that $a^2 = (-a)(-a)$, and indeed this follows immediately from the facts that $1 = (-1)(-1)$ and that $(-a) = (-1)a$, which are both given as exercises (Bartle and Sherbert 2000, p. 29). Nonetheless the proof is very close to the formal level and would be no challenge to formalize.

We can call this level of very great detail that proofs can be carried out at the "week 2 level of detail". Of course students may not see this particular argument in week 2, or at all; it is just a convenient name. We are not defining the "week 2 level of detail" here in terms of what is comprehensible to certain students– instead we give examples of basic arguments at this level of incredible detail, such as the above and also for instance basic number theoretic results (Taylor and Garnier 2014, Theorem 6.2; Silverman 2012, Lemma 7.1) or basic results from linear algebra (Axler 1997, Propositions 1.1–1.6).

As students learn the subject they won't just be passively reading proofs like this. They will also typically (and importantly) be proving these kinds of basic facts themselves, demonstrating them with arguments written out at this very explicit level of detail. The hope is that by doing this they will gain what we can call "proficiency at week 2 detail", the ability to prove simple facts like this one by chaining together these kinds of very basic steps.

A bit more will be said about how this basic level of proficiency is gained later in this section. For now we proceed onwards through the curriculum. As time passes the

arguments the students are presented with will gradually get faster, and have fewer of the details filled in. Some time later – perhaps a few months, or a term – they may encounter an argument like that in fig. I.2.

6.2.1 Interior Extremum Theorem *Let c be an interior point of the interval I at which $f: I \to \mathbb{R}$ has a relative extremum. If the derivative of f at c exists, then $f'(c) = 0$.*

Proof. We will prove the result only for the case that f has a relative maximum at c; the proof for the case of a relative minimum is similar.

If $f'(c) > 0$, then by Theorem 4.2.9 there exists a neighborhood $V \subseteq I$ of c such that

$$\frac{f(x) - f(c)}{x - c} > 0 \qquad \text{for} \quad x \in V,\ x \neq c.$$

If $x \in V$ and $x > c$, then we have

$$f(x) - f(c) = (x - c) \cdot \frac{f(x) - f(c)}{x - c} > 0.$$

But this contradicts the hypothesis that f has a relative maximum at c. Thus we cannot have $f'(c) > 0$. Similarly (how?), we cannot have $f'(c) < 0$. Therefore we must have $f'(c) = 0$. Q.E.D.

Figure I.2: Bartle and Sherbert (2000, p. 168)

This theorem presents another fundamental fact: that if a function on an interval has a "relative extremum" – a local maximum or a local minimum – at an interior point, and is differentiable there, then the derivative must be zero. This is clear by visualizing the situation, but we are doing rigorous mathematics so are not satisfied with that, and we demand a proof.

The argument given is again fairly detailed, but is slightly less explicit than the previous example: not all the details are there. It only actually covers the case where $f'(c) > 0$, showing that this cannot happen, and the task of showing that $f'(c) < 0$ cannot occur is left to the reader. If the reader has understood the argument they should have no problem seeing how this would go, or writing it out. This aspect of the proof is fundamental to the way we learn rigorous mathematics. Students will hopefully not be treating proofs like the deliverances of some oracle: lecturers will hopefully encourage them to engage with the proofs, to see if they could have proved the results themselves, to see if they can prove similar results by similar methods, and to see if they can fill in

any parts where the proof is sketchy, and check any parts of the proof they are not sure about in more detail.

The course did not start by teaching students the week 2 level of detail material just to pad the schedule. The hope is that now when they meet an argument like this which is a little bit faster, strung together out of inferences that are simple but not necessarily completely basic, they can check any inference they are not sure of by proving it in more detail, using their "proficiency at week 2 detail" that they have hopefully already attained. Thus they can sharpen their judgement of which simple (but not completely basic) inferences are valid, checking such inference whenever necessary by seeing if they can be proved.

Students won't just be seeing theorems like this however. They will also be proving these kinds of slightly higher level statements themselves, by stringing together inferences that are simple (but not necessarily completely basic). By doing so they will hopefully gain what we can call "proficiency at term 2 detail", the ability to prove these slightly higher level inferences by stringing together simple inferences, and to reliably judge the validity of simple inferences (checking whenever necessary by proving them at the week 2 level of detail).[7] Again we define the "term 2 level of detail" by giving examples, such as the above and also for instance from Bartle and Sherbert (2000, Theorem 5.2.1), Axler (1997, Proposition 2.9, Proposition 2.13) and Silverman (2012, Lemma 9.2).

As the terms go by the students are gradually exposed to more and more condensed arguments. After another year or so they might meet an argument like that in fig. I.3.

Here the students see a proof that power series can be differentiated term by term. The proof is another step up in terms of compression, in terms of relying on the intelligence of the reader. This can be seen in the first line, where the reader is expected to see

[7] A note on terminology. I find it natural to speak of *proving inferences*, as in proving them in greater detail, though strictly speaking this may be a category error: inferences are things that we draw, assess, or justify, and we normally only speak of proving statements and propositions. Nonetheless I think it is clear what is meant – replacing a given inference by a chain of intermediate inferences, which collectively constitute a proof of the conclusion from the premises – and it is a convenient and expressive idiom.

9.4.12 Differentiation Theorem *A power series can be differentiated term-by-term within the interval of convergence. In fact, if*

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \qquad \text{then} \qquad f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1} \qquad \text{for} \quad |x| < R.$$

Both series have the same radius of convergence.

Proof. Since $\lim(n^{1/n}) = 1$, the sequence $(|na_n|^{1/n})$ is bounded if and only if the sequence $(|a_n|^{1/n})$ is bounded. Moreover, it is easily seen that

$$\limsup\left(|na_n|^{1/n}\right) = \limsup\left(|a_n|^{1/n}\right).$$

Therefore, the radius of convergence of the two series is the same, so the formally differentiated series is uniformly convergent on each closed and bounded interval contained in the interval of convergence. We can then apply Theorem 9.4.4 to conclude that the formally differentiated series converges to the derivative of the given series. Q.E.D.

Figure I.3: Bartle and Sherbert (2000, p. 270)

that a certain sequence is bounded if and only if another sequence is. Also in the second sentence, the reader is expected to "easily see" that a certain equation holds. These statements are extremely plausible; and if a student has any doubt, they can check them by proving them in more detail, using the "proficiency at term 2 detail" ability they have hopefully gained. They do not need to take these statements on trust, and they do not need to guess.

Again we can talk roughly about this "year 2 level of detail", giving further examples of arguments at about this level of detail to help explicate it, again for instance from analysis (Bartle and Sherbert 2000, Theorems 9.3.2 and 10.1.3), from number theory (Silverman 2012, Theorem 42.1), and also from ring theory (Aluffi 2009, Proposition III.3.11, III.4.5) and complex analysis (Bak and Newman 2010, Proposition 3.1).

This process continues in the obvious way. As the years progress a student is exposed to gradually faster and faster arguments, arguments where gradually more and more of the details are left out and more is left to the reader's intelligence. We can pick out further levels of detail a student will encounter, in the same way as above. First, we define a "year 3 level of detail" by giving examples, now with a wider variety: from

functional analysis (Rudin 1987, Theorems 4.6–4.12), complex analysis (Conway 1978, §IV.2), measure theory (Fremlin 2010, Chapter 12), general topology (Munkres 2000, §33), algebraic topology (J. M. Lee 2000, Chapter 13), differential geometry (J. Lee 2012, Chapter 3), commutative algebra (Aluffi 2009, §V.1), representation theory (James and Liebeck 2001, Chapter 6), number theory (I. M. Niven, A. Niven, and Zuckerman 1991, §1.2), combinatorics (Szemerédi 1975, Facts 1 & 2), logic (Cori and Lascar 2000, §1.1), and category theory (Awodey 2010, Proposition 2.10). Again these may not be arguments a given student actually sees in their third year, but the level of detail is intended to be one that competent third year students will be gaining proficiency at, for both reading and writing proofs.

Detail here is not the same thing as accessibility. An argument can be very detailed but still difficult, for instance because it involves difficult concepts, or relies on difficult results, or because the result is poorly motivated and the proof strategy unexplained – or just because the argument is too long. Detail here means explicitness, and proximity to definitions, and how much the proof says of what could be said. It is the antonym of "how much is left out".

Naturally these predicates "week 2 level of detail", "year 2 level of detail" and so on will be vague: we may not be able to always determine precisely whether an argument is at the year 2 level of detail or not, just as we may not be able to decide whether a jumper is red, or perhaps orange instead. That does not undermine these predicates' validity or usefulness. Although there will be borderline cases, there will also be cases where we can in fact state with confidence that an argument is at around the year 2 or year 3 level of detail, rather than the week 2 or graduate level of detail (defined shortly). Of course an argument may not all take place at the same level of detail, so that describing the different levels of detail its parts take place at may be more appropriate than trying to shoehorn the whole argument into one category – as with a multicoloured jumper. One issue with these levels of detail that does not have such a comparison with jumpers is

that it can potentially be quite difficult to compare the detail of pieces of mathematics from very different areas, where the reasoning is of a very different style. We can mitigate that as here by giving examples from a wide range of areas when characterizing levels of detail.

We now continue in the same way, defining further levels of detail a student will encounter. It should perhaps be emphasised that this terminology of levels of detail is new terminology I am introducing, and not a standard part of mathematical discourse. There are times one might see something like it used – for instance if a mathematician presented an unconvincing argument to a colleague, and after some questioning the colleague asked them to explain it more slowly, like they were talking to a grad student. Also, concepts like these can perhaps be seen as implicitly underlying some mathematicians' talk of detail in mathematics, an example of which will be seen in section I.5 when discussing how these levels of detail can help mathematicians resolve disputes about the validity of proofs.

We will actually now pick out two different graduate levels of detail. First, we define the "graduate level of detail (explicit)" by giving examples, from functional analysis (Banach 1987, Theorem II.1), complex analysis (Conway 1978, §IV.6), measure theory (Schwartz 1954), general topology (Walker 1974, §1.1–1.16), algebraic topology (Switzer 2002, Chapter 4), differential geometry (Hirsch 1976, §1.3), algebraic geometry (Eisenbud and Harris 2000, §I.1.4), commutative algebra (Eisenbud 1995, Chapter I.2), representation theory (Fulton and Harris 1991, Lecture 4), number theory (I. M. Niven, A. Niven, and Zuckerman 1991, §5.7), combinatorics (Erdös 1947, Theorem 1), logic (Prawitz 1965, Chapters I–III), and category theory (MacLane 1998, §II.3–II.6).

The basic idea is hopefully now clear, but we can keep going and pick out a "graduate level of detail (terse)" by giving some examples: from algebraic topology (Hatcher; 2001, §2.2), differential geometry (Thurston 1997, §2.7), algebraic geometry (Hartshorne 1977, §II.3) and category theory (MacLane 1998, Chapter IX). As seen in the examples above,

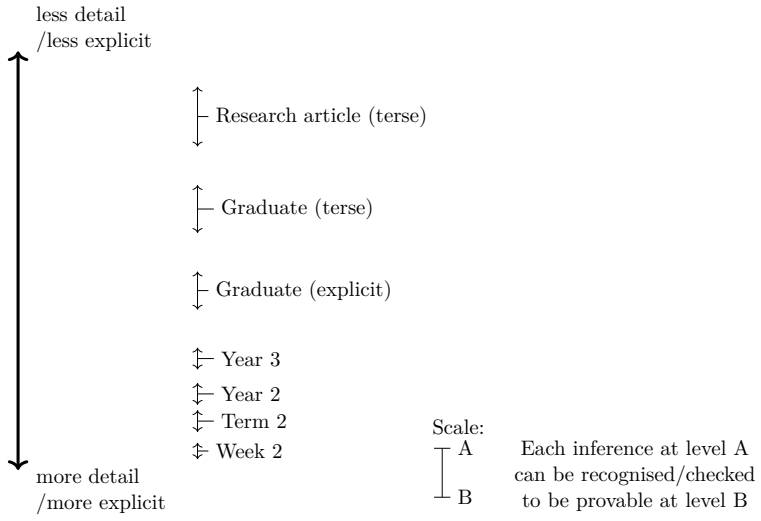


Figure I.4: Levels of detail

some research mathematics is written out at a level of detail already covered: for instance Szemerédi (1975, Facts 1 & 2) at the year 3 level of detail, and Banach (1987, Theorem II.1), Schwartz (1954), Erdös (1947, Theorem 1) and Prawitz (1965, Chapters I–III) at the graduate level of detail (explicit). However plenty of research mathematics does take place at a level of greater compression, and one could keep going and pick out levels of this, named perhaps "research article level of detail (terse)", "research article level of detail (very terse)", and maybe one or two more. A range of levels of detail we can obtain in this way is seen in fig. I.4.

To be clear, there is nothing privileged about the levels of detail listed in fig. I.4. There is a continuum of levels of detail that we could potentially pick from (perhaps idealizing somewhat, given that statements in proofs are finite objects), and nothing to mark out those in fig. I.4 as special; they are just useful examples for the purposes of this chapter.

A little more should be said about the upper reaches of mathematics, at the top of fig. I.4 and beyond. In fact it is clear that nothing too different happens as one approaches the research frontier – the gradual ascent to levels of greater and greater

compression continues. That nothing radically different is happening can be seen in the ability of professors to take the most important, beautiful or useful results in a field – once it has matured – and collate them into a textbook with proofs accessible to graduate students. In the process proofs may be simplified or altered, but there is never any great obstacle to writing out what was once research level mathematics at a level of detail that graduate students can follow. One can see this in for instance the titles in the Springer Graduate Texts in Mathematics series.

A particularly extreme example of this is given by the case of Perelman's proof of the Poincaré conjecture. This was one of the most major conjectures in mathematics, and the subject of a Clay Millennium Prize. In 2002 and 2003 Perelman uploaded three papers containing a claimed proof (Perelman 2002; 2003b,a). The papers were written at a very high level, containing mathematics sketchy enough that despite only totally 70 pages, it took teams of mathematicians 3 years – working in correspondence with Perelman – to verify the argument as correct. Perelman was then offered a Field's medal, and subsequently a Millennium Prize, both of which he declined. Since 2006 more detailed expositions of his argument have been produced, such as Morgan and Tian (2007), which is a textbook intended to be accessible to graduate students. Indeed it looks to be at about the graduate level of detail (explicit) or graduate level of detail (terse). It comes in at 521 pages – about 8 times longer than Perelman's original papers. This is quite an increase, but even so it shows that there is not too dramatic a leap in terms of compression from maths at a level graduate students can understand to some of the most concise mathematics acceptable as a proof.

Now to say a little more about the lower end of fig. I.4. First, a potential issue is that some areas of maths are generally only studied at a high level, because they have substantial mathematical prerequisites or typically involve subtle or complex arguments. This is the case with harmonic analysis, the theory of functions of several complex variables, and modern algebraic geometry, amongst other areas. This presents a potential

problem with regards to the lower levels of detail in fig. I.4, since for instance no-one does algebraic geometry in week 2 of their degree. In some cases this is unproblematic since expanding an argument in great detail will lead to basic inferences like those found in other areas – for instance combinatorics or analysis, in the case of harmonic analysis. In other cases reasoning used is more *sui generis*. Nonetheless in these cases I think it still makes sense to talk about what an inference carried out at say the term 2 level of detail would look like. In fact we sometimes see this happen: when a new manner of argument is introduced, even to advanced students, a few very explicit examples are often given of how it works. This is so that students have a sense of what is underlying more complex arguments, and know what to fall back on if they ever find more complex arguments hard to follow or produce. An instance of this from differential topology – a fairly advanced subject, done in full generality – is seen in J. Lee (2012, Proposition 2.4). Here Lee is giving an example of how to use the smooth charts on a manifold to prove local facts about them, and it is written up in great detail to make clear to students how this works, at around the week 2 or term 2 level of detail. One could do the same for other advanced subjects, for instance writing out the arguments of Eisenbud and Harris (2000, §I.1.4) from algebraic geometry mentioned above at the term 2 level of detail.

A second issue about levels of greater detail is one that was rather passed over in the discussion above: what takes place in the initial stages of learning rigorous mathematics, before the ascent up the levels of compression can start. To begin with, students will be taught how the basics of proof work, and what to do with the logical vocabulary $\wedge, \exists, \neg, \forall$ and so on, by a combination of examples and informal descriptions of what is going on and why. For instance they will (hopefully) learn by seeing and working through examples that to prove a statement $\forall n\, \phi(n)$ about all numbers, one can take an unknown number n, prove that $\phi(n)$ holds without assuming anything special about n, and then deduce that indeed $\forall n\, \phi(n)$ holds – essentially the $\forall$-introduction rule from natural deduction. The logical workings of proof can be stated in simple, clear, precise

form – in terms of the hypotheses active at each stage, and how to introduce and exploit them – and when students can grasp how this works from examples it is not such a great surprise. Anyway not all students do manage to learn the rules from examples, and how to best teach the logic of proof is much discussed in the mathematics education literature (see for instance Epp 2003; 2009). As well as the logical vocabulary students will learn the basics of set theory, including how to determine if two sets are equal via extensionality (whether this is an axiom or an inference rule is not important to the students, and the distinction may not be clear to them), and they will be shown various acceptable ways of forming new sets – power sets, cartesian products and so on.

There is a further basic aspect of proof that students are expected to infer from examples, and this is the ability to prove results by describing algorithms or procedures to achieve some desired mathematical goal, with Euclid's algorithm being an early example students often encounter (Silverman 2012, pp. 33–34), and further examples coming in linear algebra (Axler 1997, Proposition 2.6, Proposition 2.7; Artin 1991, pp. 14–15) and other areas. Though some such proofs can be rephrased as arguments by induction, sometimes they may implicitly require the definition of functions by recursion, and this is not usually something early undergraduates will be in a position to justify formally – the set theoretic treatment of recursion is typically taught later on in a more general form that applies to recursions on all ordinals (such as in Jech 2006, Theorem 2.15). In fact this does not present any sort of problem, and is not so different to the cases above where students grasp principles from examples; in this case the general principle implicitly underlying these kinds of recursive arguments is the axiom of dependent choice. This states that if X is a set and R is a binary relation on X such that for all $x \in X$ there is $y \in X$ with xRy, then for all $x \in X$ there is a sequence $(x_0, x_1, \ldots)$ of elements of X where $x_0 = x$ and for all i, $x_i R x_{i+1}$. That this is a statement rather than an inference rule, and not a basic axiom of set theory (it is deduced from the axiom of choice) is not important here. It is intuitive, and can be stated clearly, simply and precisely, and it is

not a surprise that students can grasp from examples what kinds of arguments are in line with it.

The basic axioms governing sets are exempted from the general requirement that inferences be justifiable by proofs in greater detail. One can rightly assert this without actually deciding on which the basic axioms are; for instance it does not matter whether one regards the statement that function sets B^A exist as a basic axiom, or as justified by an argument that appeals to more basic axioms (union, separation, pair set and power set perhaps). Whatever the basic axioms are, they need not be justified by a proof, and other basic properties of sets are justified in terms of them (perhaps out of sight of students). There has been some discussion of exactly what means of set formation are allowed in mathematics – in particular, whether the unrestricted comprehension principle of naive set theory is used – and this issue will be addressed in section I.7.

With these points addressed, we have the essentials of how rigour is learnt. We can pick out different levels of detail maths is done at by giving examples, and students proceed upwards through these levels of detail as described above: once they have gained proficiency at a certain level of detail they are in a position to engage with more concise arguments, with a tutored sense of what more compressed inferences are valid – tutored by their experience at proving such inferences. If they are ever unsure, they can use their existing proficiency to check a more compressed inference and see if it can in fact be justified by a proof; and if so, they can ask why they were suspicious about it, and consider how to adjust their instincts to recognize such inferences as valid in future. It is essential to the normal process of learning rigorous mathematics that students are in a position to check inferences for themselves in this way, rather than just being presented with high level arguments they are intended to imitate. This is the most significant difference between mathematics as taught rigorously, and mathematics as taught in a physics degree (for instance). As discussed above there are also cases early on where students are expected to infer general principles from examples. There are

only a few such cases though, and the reasoning each general principle encompasses can be characterized simply and precisely.

As mentioned in section I.1, there are various places where this account could be expanded on. This could be a task for further investigations on the subject. For instance, one might seek a better understanding of exactly how students "adjust their instincts" to recognize a wider variety of inferences as valid, having seen particular cases to be provable in greater detail; or how, and with what degree of success, they "infer general principles from examples" in the early stages. These questions will be put to one side in this thesis however.

One limitation of the above that will be addressed later is that much of what has been said only really makes sense for sentential arguments, rather than for instance arguments involving pictures; what rigour requires of pictorial arguments in particular is considered in chapter III.

4 The concept of rigour

Implicit in the process of learning rigorous maths described above is that each time a student is trying to master a new level of greater compression, it is constitutive of inferences being at that level of compression that they be provable at a previous level of detail, a level the student is already comfortable at – so that there are no leaps in the process of learning rigorous mathematics where a student is unable to check inferences for themselves (apart from when grasping certain basic principles). Indeed if inferences at the level of greater compression didn't need to be provable in more detail, then "checking" them by seeing which inferences can be justified with a proof would be a mistake. As discussed above, the basic axioms of set theory are exempted (whatever exactly they are taken to be), and are intended to be accepted by students without argument, though possibly with the assurance that they are "obvious".

Similarly, it is constitutive of rigour that nontrivial inferences can potentially be proved at a level of greater detail, so that if there is ever any unclarity – or disagreement – about the correctness of an inference it can be resolved by seeking a proof, or requesting one. Again proving in more detail here means an appreciable step up in detail, going up a notch in terms of levels of detail that we can pick out. This is one useful mechanism in mathematics for resolving disputes about the validity of proofs, as discussed in section I.5. For the purpose of resolving disputes, this requirement is of less practical importance at levels of very great detail, as mathematicians may agree immediately about sufficiently basic inferences; but equally, for these simpler inferences it is generally more obvious that they can be proved in greater detail, and how such a proof would go, so the requirement is no unnecessary burden. Also, the ability to gain greater clarity about the correctness of inferences at all levels by seeking a more detailed proof is very important for the purposes of learning mathematics, as mentioned above. Again, the basic axioms of set theory are exempted from the requirement of being justifiable in greater detail: their correctness is not up for debate (within mathematics) and is supposed as a precondition for the mathematical enterprise to get going.

This is, I think, the key feature of rigorous mathematics: that there is a range of levels of detail that it can take place at, where inferences at a more compressed level can necessarily be proved at an appreciably greater level of detail. This requirement that inferences be provable in more detail is why the examples discussed in section I.2 were not rigorous: the manipulations of infinitesimals and infinite series discussed could not be justified in greater detail, and could not be regarded as basic rigorous rules in themselves since it was not clear how to demarcate what reasoning was acceptable.[8] Similarly brute intuitions such as for the truth of the Jordan curve theorem are not in themselves suitable in proofs unless they can be backed up with more detailed arguments.

[8]Reasoning with infinitesimals was put on a firm footing by Robinson (1996) in terms of the logical concept of non standard model, and Nelson (1977) showed how to give a rigorous axiomatization for the approach. One can also use a paraconsistent approach to embrace the contradictory nature of infinitesimals, as seen in Brown and Priest (2004) and Sweeney (2014).

The necessity that inferences be provable in greater detail applies to levels of detail that students (and mathematicians) reach after progressing onwards from the basic level at which the subject is first taught – assumed above to be the week 2 level of detail. However not all students do manage to directly grasp how proof works at this level of detail, and some need more explicit demonstration of the rules that proof is implicitly following. There are courses and textbooks which provide this, such as Velleman (2006), which teaches how proof works essentially by teaching how to prove statements using the natural deduction rules: here the premises being used are explicitly tracked and calculated with, according to the rules governing the various bits of logical vocabulary. We can call this most basic, most explicit level of detail the "intro to proof level of detail". Students can use this as a stepping stone to gain comprehension of how basic arguments at the week 2 level of detail work, and implicit in this (as above with later levels of detail) is that inferences at the week 2 level of detail be provable at the intro to proof level of detail – otherwise gaining a grasp of how arguments work at the intro to proof level would be misleading as to what is going on at the week 2 level. Thus we can extend the above argument that inferences need to be provable in greater detail all the way down until we reach the intro to proof level of detail, where arguments explicitly use the natural deduction rules.

A minor caveat to this ability to prove in greater detail is that some arguments may require rephrasing when proving in more explicit formal terms – for instance one would often justify the Euclidean algorithm as an informal recursive process of repeated division with remainder, described for integers $a, b > 0$ perhaps by saying:

$$\text{Write } a = q_0 b + r_0 \text{ with } 0 \leqslant r_0 < b,$$
$$\text{write } b = q_1 r_0 + r_1 \text{ with } 0 \leqslant r_1 < r_0,$$
$$\text{write } r_0 = q_2 r_1 + r_2 \text{ with } 0 \leqslant r_2 < r_1,$$
$$\text{and so on, until we reach } r_n = 0$$

which if proved more explicitly would be transformed into some sort of formal recursive

definition (of the form justified by the axiom of dependent choice – though there is no choice here – as discussed in section I.3). In these kinds of cases one is replacing a part of a proof with a similar more detailed version, rather than literally filling in inferences in greater detail. However these kinds of cases are the exception rather than the rule, and do not make any essential difference to any of the discussion below, so we will generally include them under "proving in greater detail" (speaking a little loosely).

Now although the account above has focused on the ability to prove inferences in greater detail, it should be emphasised that one does not generally have to see how to prove an inference in greater detail to understand it, or accept it. For instance, to someone familiar with the notion of homeomorphism it probably feels obvious that the sphere $\{x \in \mathbb{R}^3 : \|x\| = 1\}$ is homeomorphic to the cube $\{x \in \mathbb{R}^3 : \max_i |x_i| = 1\}$, but sitting down and trying to write out a proof of this could well take a while. Such examples are not limited to topology. In logic – a subject where perhaps one would expect "intuition" would play less of a role – it might well feel obvious that substituting term t for variable x in a formula ϕ, when x is not free in ϕ, will just return ϕ, but again proving this in detail would take a bit of work (though probably less insight than the previous example). These kinds of higher level judgements about inferences – without a proof in mind – are an essential part of mathematics. Nonetheless it is important in rigorous mathematics that the option of proving inferences in more detail is there, to aid in gaining a firm grasp of any new concepts, and in guiding and sharpening one's judgement in any difficult cases – not all homeomorphic spaces are as obviously homeomorphic as the two above.

Also, when checking a proof one does not always necessarily actually check that every inference in is valid. Indeed if an unsurprising claim in a proof is supported by reasoning that looks like the right kind of thing, and the right amount of effort, then an experienced mathematician may pass on without checking every single detail. This is seen in interviews conducted by Andersen (2017) with mathematics referees, and also

mentioned by Thurston (1997, p. 32). It appears this can be a fairly good guide to the overall correctness of results, though numerous commentators have remarked on the unreliability of the mathematical literature and the pervasiveness of errors in proofs, to which this manner of refereeing may be a contributing factor (Jaffe and Quinn 1993; Thurston 1994, p. 33; Nathanson 2008; Grcar 2013, pp. 421–422).

5 Disagreements about validity

It is traditional when studying deduction to think in terms of a single sharply defined notion of validity, that every inference either has or lacks. I think the above analysis of rigour rather tells against this conception.

Indeed having picked out various levels of detail that mathematical inferences can take place at, we can introduce a cumulative hierarchy of validity predicates – "valid at the week 2 level of detail", "valid at the term 2 level of detail", and so on, where for instance "valid at the term 2 level of detail" means an inference either at the term 2 level of detail, or at a level of greater detail. One can keep ascending in this way, defining validity predicates which allow larger and larger, more and more compressed inferences, inferences that are increasingly challenging for even an experienced mathematician to follow. At some point one will reach inferences compressed enough that they are well beyond the bounds of what mathematicians consider to be valid. However there appears to be no natural place on the continuum of levels of detail to draw a line, and say *this* is the limit of validity: that inferences at least that detailed are valid, while those less detailed are not. Asserting that there is a precise such limit seems to just be philosophical dogma, unsupported by the facts of the practice. Suppose for instance that a preprint of an article is uploaded to the arXiv, and read by two mathematicians experienced in the field – one of whom concludes that it is perfectly rigorous, the other that the proof of a certain lemma is too sketchy and incomplete. Who are we to say as philosophers

that one is definitely right, and the other definitely wrong? I think it is more plausible to say that judgements of validity are – to some extent – both vague and subjective.

Some philosophers are aghast at this suggestion, and it is worth explaining why, given the account of rigour above, it is not as damaging a claim as might be thought. In as much as there is disagreement about the validity of a certain inference, the framework of rigour provides a mechanism for resolving it.

That mathematicians are good at resolving disagreements about validity – that if a mathematician believes a proof is valid, they are generally able to convince others of this, or become convinced themselves of a flaw in it – has been noted by various authors as a fact that deserves explanation, and is one of the main motivations of Azzouni's derivation indicator view of proof is to attempt to provide this (Azzouni 2004, pp. 83–84; Antonutti Marfori 2010, pp. 267, 270–271). The idea behind Azzouni's account is that the informal proofs mathematicians write serve to indicate formal derivations. This has met with criticism, with Tanswell (2015) pointing out that attempted proofs may have many different attempted formalizations, which poses a problem since Azzouni wants to characterize validity of the informal proof in terms of success of the indicated formal derivation. Additionally as seen in section I.3, mathematics generally proceeds at a much greater level of compression than is found in formal proofs, and (as mentioned in section I.1) most mathematicians have no experience of or interest in the activity of formalization, so Azzouni's account is rather far removed from how mathematicians typically engage with proofs in practice.

Instead of hoping for an explanation in terms of formal derivations, it is more promising to look to the process of proving in more detail itself. Indeed if a mathematician ever puts forward a purported proof in which an inference is not convincing, then a more detailed justification for that inference can be requested – an appreciable step up to a level of greater detail, perhaps from the research article level of detail (terse) to

the graduate level of detail (terse), as discussed in section I.3.[9] At this level of greater detail inferences are more transparent and judgements of validity are more reliable, and this may already serve to resolve the controversy – with the new argument being acceptable, or an obvious error in it discovered. If not, and an inference in this more detailed argument is still controversial, a more detailed justification can be asked of it in turn, taking us to a still greater level of detail at which errors will be even more obvious – perhaps we now reach the third year level of detail. In principle this process will terminate when one reaches the level of complete formalization, though in practice if both sides are of sound mind and proceeding in good faith then the controversy will be resolved well before that.[10]

Thus if one believes a proof to be rigorous, in as much as this belief is correct one can always (in principle, and usually in practice) fill in the details of any inferences that are felt to be sketchy, to increase the level of detail to one which is found acceptable. In the imagined case considered initially of two mathematicians disagreeing about an arXiv preprint – with one finding that a particular lemma was argued for too briefly – the dispute would normally be resolvable in this way, bringing the proof into a form acceptable to everyone.

Thus though in my view there may be some subjectivity and vagueness to where exactly the limits of rigour are drawn, much more significant is the strong form of objectivity afforded by rigour, in which there are always robust reasons available to win over an objector – provided one's assessment of a proof as valid is justified.

In fact this dispute resolution mechanism is essentially that described in the Princeton Companion to Mathematics:

> [T]he fact that arguments can in principle be formalized provides a very

[9] Though this terminology of levels of detail is new, the process is not.

[10] The resolution of disagreement in this manner is an example of the idea from argumentation theory that a debate goes down to the level of detail that will satisfy both parties (using the apt words of an anonymous referee). For more on argumentation theory and mathematics, see Aberdein and Dove (2013).

> valuable underpinning for the edifice of mathematics, because it gives a way of resolving disputes. If a mathematician produces an argument that is strangely unconvincing, then the best way to see whether it is correct is to ask him or her to explain it more formally and in greater detail. This will usually either expose a mistake or make it clearer why the argument works. (Gowers, Barrow-Green, and Leader 2008, p.74)

Sections I.3 and I.4 can be seen as a clarification and elaboration of the process outlined in this quote. As an explanation of how disputes in mathematics can be resolved, this seems to be both more straightforward and more plausible than Azzouni's account, and better grounded in mathematical practice. However there may well be more to be said about the reality of disputes over validity in mathematics, and how well this simple account fits it.

6 Naive set theory?

The final topic of the chapter is how this account of rigour impacts on the question of formalizability. Before the discussion of this in section I.7, we take a brief detour from the main course of the chapter, to address a sceptical view about the basic principles used in mathematics: the idea that mathematicians should be viewed not as working in a system like ZFC(U), but in naive set theory, with its axiom scheme of unrestricted comprehension. This has been suggested by Leitgeb (2009), and some mathematicians have made similar claims about their own understanding (Jones 1998, p. 205; Aluffi 2009, p. 1). Indeed "Naive set theory" is actually the title of a set theory textbook by Halmos (Halmos 2011). If mathematicians are best regarded as using unrestricted comprehension, this would make the claim that mathematical proofs can be formalized trivial, since set theory with unrestricted comprehension is inconsistent and so any argument can be immediately formalized in it (with every inference justified by a set theoretic paradox).

The key step when considering this possibility is to distinguish different senses of the term "naive set theory". Certainly most mathematicians do not know what the axioms of ZFC are, but they do have a solid grasp of how to legitimately form new sets: by taking unions, subsets, power sets, Cartesian products, function sets, equivalence classes, and so on. This understanding may be "naive" in the sense that it is not accompanied by explicit awareness of how these operations are justified in terms of the basic axioms – but that is totally different to "naive" set theory in the logicians' sense, in which the central principle is that of unrestricted comprehension, the scheme

$$\exists y\ (x \in y \Leftrightarrow \phi(x))$$

for all formulae ϕ in which y does not occur free.[11] Indeed the above set forming operations are not generally justified in any more direct terms by unrestricted comprehension than by ZFC; to form Cartesian products for instance, one still needs to go through the rigmarole of defining what ordered pairs are and what a family of objects is, and the availability of unrestricted comprehension does not significantly simplify this. Moreover, there is no evidence of mathematicians making essential appeals to unrestricted comprehension, and this being accepted as valid. In the normal course of mathematics, all classes one would like to be sets are easily seen to be set sized using the standard set forming operations. In category theory, where size issues are encountered, the axiom of universes was introduced specifically so that they could be dealt with in a rigorous way. There are occasional instances where classes are manipulated as though they were sets, for instance in the definition of the Grothendieck group as a quotient of the set of isomorphism classes of finitely generated modules over a ring R; but in this case there is nothing genuinely troubling going on since one can easily define a set of representa-

[11] The set theory in Halmos's textbook is naive in an even weaker sense. Halmos does in fact state the usual basic axioms of set theory (with no mention of unrestricted comprehension), and he uses them to derive various set theoretic operations and facts, but he says the approach is naive in that "the language and notation are those of ordinary informal (but formalizable) mathematics".

tives of the isomorphism classes instead (the quotient modules $\frac{R^n}{M}$ of powers of R, with isomorphic quotients identified), or one can appeal to the axiom of universes.

Moreover there are genuine mathematical cases where the distinction between sets and classes is crucial, and the use of unrestricted comprehension would be disastrous. For instance the general adjoint functor states that if $G : D \to C$ is a functor with D complete and locally small, then G has a left adjoint iff it preserves all limits and satisfies the solution set condition (MacLane 1998, p. 121). The solution set condition states that a set of morphisms with a certain property exists, and in this case the fact that this be a set rather than a class is key, as there is always a class of morphisms with the required property. In the presence of unrestricted comprehension, the general adjoint functor theorem would become the claim that any limit preserving functor whose domain is complete and locally small has a left adjoint, which is false in general. The issue of which functors have adjoints is not some category theoretic curiosity – it arises naturally in various areas of mathematics including algebra and topology.

Thus the way mathematicians form sets may be naive in the sense that it need not be founded in explicit knowledge of the basic principles, but there is no indication that it is naive in the sense of relying on unrestricted comprehension. If it did, signs of this ought not to be too hard to find.

7 Formalizability

Now to the question of formalizability itself. As advertised previously, it will be argued here that valid rigorous proofs are formalizable, in principle, though what this means requires clarification. The prospects for feasible formalization will also be touched on.

For the purposes of this section we will introduce the concept of "deductive grounding" between levels of detail. If L, L' are levels of detail, then we say that L' is *deductively grounded* in L if every inference valid at level of detail L' is provable at level of

detail L.[12] This concept of grounding is justificatory rather than metaphysical, and does not share all the properties of standard notions of metaphysical grounding (for instance it is reflexive). This concept of grounding uses a notion of provability, and the kind of modality employed here needs to be spelt out before we know what deductive grounding amounts to. The notion of provability we will use is one of "in principle" provability, abstracting from limitations in terms of time or other resources (and thus perhaps abstracting away from the limitations of our own physical universe). One could otherwise use a notion of what is actually feasibly provable, given the physical and biological constraints on us, and obtain thereby a notion of "feasible deductive grounding".

We can relate this notion of deductive grounding to a notion of "in principle formalizability". Indeed, we can fix some standard system of first order natural deduction (for instance that of Prawitz 1965, Chapter I), and we take a formal derivation to be a complete derivation in this system in the language of set theory with all premises amongst the axioms of ZFCU. We then define the "formal level of detail" to consist just of these formal derivations. We characterize a proof as being *formalizable in principle* if every inference in it is provable at the formal level of detail. Thus the claim that every proof at a level of detail L is formalizable in principle is just the claim that L is deductively grounded in the formal level. This is a weak notion of formalizability in principle, and for instance if the Riemann hypothesis is a theorem of ZFC then the one line proof of the Riemann hypothesis from no premises is formalizable in principle by this definition. Nonetheless it is one possible sense of formalizability in principle; the question of what we do and should mean by formalizability will be returned to later in this section.

The key property of this notion of in principle provability is that it satisfies a version of the converse Buridan formula. In general the converse Buridan formula is (the scheme

[12] As was noted in section I.4, some inferences at level L' may be part of an informal section of a proof which as a whole needs to be replaced by a more detailed and formal version at level L; this was illustrated with the example of the Euclidean algorithm. This caveat makes no real difference to what follows.

of formulae) of the form

$$\forall x \, \Diamond(\phi(x)) \Rightarrow \Diamond(\forall x \, \phi(x)).$$

It is of the same general form as the converse Barcan formula, though with an existential instead of universal quantifier (Konyndyk 1986, p. 94). This converse Buridan formula is not typically valid. For instance if I have a well stocked fridge, it may be the case that for every item in the fridge I can have that item as part of my dinner, but that it is impossible for me to have every item in the fridge as part of my dinner. Nonetheless in cases where one abstracts from resource constraints it can be valid. In particular if we have a finite set S of inferences, and write Prov to indicate that we have obtained a proof of $s \in S$, then we do have that

$$\forall s \in S \, \Diamond(\mathrm{Prov}(s)) \Rightarrow \Diamond(\forall s \in S \, \mathrm{Prov}(s))$$

since if each element of S is provable, then – given sufficient time – it will be possible to obtain coeval proofs of every element of S.

It follows that if level of detail L' is deductively grounded in level of detail L, and we have a proof p of result s valid at level of detail L', then one can in principle obtain a proof of the s at level of detail L. Indeed every inference in p is provable at level of detail L, so that (as just discussed) it is possible to obtain a simultaneous proof of every inference in p; concatenating these then gives a proof of s at level of detail L, as claimed.

Thus we can obtain that the notion of deductive grounding is transitive. Indeed suppose we have levels of detail L, L', L'' with L' deductively grounded in L and L'' deductively grounded in L'. Let s be an inference valid at level of detail L''. Then by the definition of deductive grounding, we can in principle obtain a proof of s at level of detail L'; but then as just discussed, given such a proof one can in principle obtain a

proof of s at level of detail L. Thus s is indeed provable at level of detail L.[13]

Now we will use this to argue that for the levels of detail that a student moves through on their way to mastering research level mathematics, each more compressed level is deductively grounded in its more detailed predecessors.

First we will give an informal argument for this. Recall that as discussed in section I.4, it is crucial for rigour that valid mathematical inferences be provable in greater detail (unless they are already basic). This is an essential part of how rigorous mathematics is learnt, and of how validity is reliably judged – since for inferences that are not immediately convincing, one can always clarify the situation by seeking a proof in more detail. In both cases proving in more detail means an appreciable step up in detail, going up a notch in terms of levels of detail that we can pick out. This is an important mechanism for resolving disputes in mathematics, as seen in section I.5.

This requirement that inferences be provable in greater detail does not immediately imply that (for instance) the year 3 level of detail is deductively grounded in the year 2 level of detail, since perhaps one could point instead to some intermediate level of detail L used as a stepping stone to reach the year 3 level of detail. But it would not be possible to pick out more than a small finite number of levels of detail between the year 2 level of detail and the year 3 level of detail that we could actually distinguish – with each being an appreciable step up in compression from the last – and that a student would have time to move through, using each one as a stepping stone to the next, in the course of ascending from the year 2 to the year 3 level of detail. If these were listed in order as $L_1, L_2, \ldots L_n$, with L_1 deductively grounded in the year 2 level of detail, L_{i+1} deductively grounded in L_i for each i and the year 3 level of detail deductively grounded in L_n, then it follows immediately that the year 3 level of detail is deductively grounded in the year 2 level of detail, by transitivity. Similarly we obtain that the term 2 level of detail is deductively grounded in the week 2 level of detail, that the graduate level of

[13]This argument could be carried out more formally, and implicitly uses the rule $\Diamond\Diamond\phi \Rightarrow \Diamond\phi$ of S4, which holds for the kind of metaphysical possibility being employed.

detail (terse) is deductively grounded in the graduate level of detail (explicit), and so on. Thus we obtain by transitivity that all such levels of detail are deductively grounded in the week 2 level of detail.

Now we give a more formal argument. We will consider a student gradually moving through an education in mathematics, where at each time t there is the level of detail m_t of mathematics that they have mastered so far, and the level of detail l_t that they are learning at that point. We will assume that the collection of times t making up this period of education forms a complete ordered set, which we denote by I. I may or may not have endpoints – an initial point, and/or a final point. If L and L' are levels of detail we will write $L \leqslant L'$ to denote that L' consists of mathematical inferences at least as compressed as those at level L. We will assume that the function $t \mapsto m_t$ is monotonic, i.e. that if $t' \geqslant t$ then $m_{t'} \geqslant m_t$. For there to be no magical jumps in this process of learning, it needs to be the case that for all times t not initial in I, there is some $t' < t$ such that $m_t \leqslant l_{t'}$, so that the level of detail mastered at time t was actually learnt at some previous point in time. We also need that for all times t not final in I, there is some $t' > t$ such that $m_{t'} \leqslant l_t$, so that there is no magical jump after time t where at all subsequent times, a level of detail has been mastered that is greater than that which was being learnt at time t. Finally, as discussed above, it is constitutive of learning rigorous mathematics that at each stage, the level of detail one is learning is deductively grounded in a level of detail one has already mastered, i.e. that l_t is deductively grounded in m_t.

Now for the argument. For each t there is some $t' < t$ such that $m_t \leqslant l_{t'}$, but $l_{t'}$ is deductively grounded in $m_{t'}$, so that by transitivity m_t is also deductively grounded in $m_{t'}$ (call this backwards grounding). Also, for each t, there is $t' > t$ such that $m_{t'} \leqslant l_t$, so that $m_{t'}$ is deductively grounded in m_t (call this forwards grounding). Now suppose for contradiction that there is $s > t$ such that m_s is not deductively grounded in m_t. We

let r be the infimum of

$$\{s \mid m_s \text{ not deductively grounded in } m_t\}$$

(a set which is bounded below by t). By forwards grounding, there is $t' > t$ such that $m_{t'}$ is deductively grounded in m_t, so that by monotonicity of m, we have that $r \geqslant t' > t$. Thus r is not initial in I, and so by backwards grounding for r, there is $t' < r$ such that m_r is deductively grounded in $m_{t'}$. But by the definition of r, we must have that $m_{t'}$ is deductively grounded in m_t, and thus by transitivity that m_r is deductively grounded in m_t. But then forwards grounding for r gives that there is $t' > r$ such that $m_{t'}$ is grounded in m_r, and thus in m_t; then by monotonicity of m, if $r \leqslant t'' \leqslant t'$ then $m_{t''}$ is deductively grounded in m_t, so that t' is actually a lower bound for

$$\{s \mid m_s \text{ not deductively grounded in } m_t\},$$

so that r is not the infimum of this set, giving the required contradiction. Thus from the assumptions we have made, it follows that if $s > t$, then m_s is deductively grounded in m_t.

This we obtain that all levels of detail that can be mastered rigorously, from a process starting at the week 2 level of detail, are deductively grounded in the week 2 level of detail. Then it is clear by inspection that arguments at the week 2 level of detail can be written out as formal derivations, as was noted in section I.3, and their premises are all amongst the basic axioms of set theory, so that the week 2 level of detail is deductively grounded in the formal level. Otherwise, as discussed in section I.4, one can define an "intro to proof level of detail", in which the natural deduction rules are explicitly used, and which some students use as a stepping stone to master the week 2 level of detail; it follows (as above) that the week 2 level of detail is deductively grounded in the intro to proof level of detail, whose proofs are trivially formalizable – they are already

natural deductions of a slightly nonstandard kind. Either way we obtain that the week 2 level of detail is deductively grounded in the formal level, and thus that all the levels of detail that can be mastered rigorously are deductively grounded in the formal level. In other words, all rigorous mathematics is, in principle, formalizable. This is not a mere empirical fact obtained by looking at examples of mathematical arguments (except perhaps for the step from the formal level to the intro to proof or week 2 level), but a direct consequence of the norm of rigour in mathematics as enunciated here in terms of levels of detail.

We can also note the answer to a concern about formalizability noted in section I.1: if rigour requires proofs to be formalizable, how are mathematicians so good at judging this? The answer is that mathematicians are not directly judging formalizability, but are instead judging the rigour of inferences (as discussed in sections I.3 and I.4), and that formalizability is obtained as a consequence of rigour.

In the literature the main worry about this kind of in principle formalizability is raised by Rav (1999), reiterated by Weir (2016). Rav considers a situation where one has an inference $A \to B$ in a proof, and after some thought fills this in with intermediate inferences to obtain $A \to A_1, A_1 \to A_2, \ldots, A_n \to B$. Perhaps one is then questioned by a student or non specialist as to why A_1 follows from A, and comes up with a new interpolation $A \to A', A' \to A_1$. Rav claims we can give no "theoretical" reason why this process of adding interpolations ought to ever terminate (Rav 1999, pp. 14–15).

The basic problem with this description is the lack of any attempt to characterize the form the interpolating inferences must take. One cannot just write in any intermediate inferences that suit one's fancy, whether justifying an inference to oneself or to a sceptic. For instance suppose we are trying to argue that the fact there are infinitely many primes (IP) is a consequence of the fundamental theorem of arithmetic (FTA), and we write RH for the Riemann hypothesis and ST for Szemerédi's theorem. It is clearly nonsense

to try to justify FTA $\to$ IP by filling in intermediate inferences of the form

$$\text{FTA} \to \text{RH},\ \text{RH} \to \text{ST},\ \text{ST} \to \text{RH},\ \text{RH} \to \text{IP}.$$

These may be valid as material implications (in the presence of the axioms of ZFC), but they are totally useless as intermediate inferences for us; and to explain why they are the wrong kind of intermediate inferences, we have to start putting inferences on some sort of scale of plausibility, or simplicity, or evidentness, and require that adding intermediate inferences take us in the increasing direction on this scale. This is the first step towards thinking of mathematics in terms of levels of detail, as in sections I.3 and I.4, and to the requirement – implicit in Rav's description – that nontrivial inferences be provable in greater detail. We can then argue as above that rigour implies formalizability, giving the dissolution of Rav's worry.

The argument given above concerned in principle formalizability, but it is also possible to say a little from this perspective about how practical formalization might be. Indeed we can use the above framework to address a weaker version of Rav's worry put forward by Pelc (2009), who accepts that the process of filling in with intermediate inferences will necessarily terminate, but questions whether we have any reason to believe this process will result in a formal derivation of feasible length (given some initial proof of reasonable size). Pelc defines a vast number M in terms of various fundamental constants, large enough so that there is no possibility of us ever (in practice) constructing a formal derivation of this length. His number M is at least the number of particles in the universe divided by the Planck time (in seconds), and thus at least 10^{120} on standard estimates; so a special case of Pelc's worry is whether when formalizing a proof of reasonable length, we have any reason to believe the resulting formal derivation will be less than 10^{120} symbols.

The framework of levels of detail developed above can also be useful when addressing

this kind of worry. Indeed one can get from the week 2 level of detail to the research article level of detail (terse) in a small number of steps up: via the year 2 level of detail, the graduate level of detail (explicit), and the research article level of detail (explicit). For each of these steps one can estimate by considering examples what kind of factor increase in length one generally obtains, when writing out an inference from the more compressed level at the more detailed level. Though it would have to be confirmed by more careful investigation, my belief from considering examples is that this factor is not generally too large, with a factor of less than 5 being common, a factor of 10 being fairly rare and a factor of 20 being very rare (as a proportion of inferences). This is supported by the example of Perelman's proof of the Poincaré conjecture discussed in section I.3, where an extremely high level argument was written out at around the graduate level of detail with a factor of increase in length of 8. Thus though one may obtain exponential growth in proof length as one fills in details in a proof, to bring it down to the week 2 level of detail, the base and exponent are both fairly small: the former being the factor of increase with each step to greater detail, the latter being the number of such steps (4 in the case above). Even with a factor of 20 at each stage, and another crude upper bound factor of 20 to get to the formal level from the week 2 level, this gives us an overall factor of increased length of at most $20^5 = 3,200,000$ to formalize an argument at the research article level of detail (terse). This is vastly less than the factor Pelc is concerned might be necessary. Though this could undoubtedly result in very unwieldy proofs (if this crude upper bound was attained), they would still be within the bounds of feasibility, as usually conceived – requiring perhaps a few gigabytes or tens of gigabytes of space to store on a hard drive.

Even if one can fill out the above sketched argument to make a convincing case that all rigorous proofs of reasonable length will be feasibly formalizable, this does not meet one major kind of objection to formalizability: namely, that the process of formalization so dramatically changes a proof that the formal derivation that results cannot rightly

be regarded as the "same proof" as the original, and thus should not be regarded as a formalization of it. Larvor writes of the "violence or essential loss" that can result from formalization (Larvor 2012, p. 717). A related question concerns what practical relevance formalization has, or could have, to mathematics; it played only an indirect role in the account of rigour from sections I.3 and I.4, and turned out to be inessential to the process of resolving disputes in mathematics described in section I.5. These questions will have to await possible consideration in future work.

8 Further questions

Before continuing to examine case studies, a few issues left unresolved about rigorous proof should be mentioned.

First, to say a bit about a topic that has been neglected: understanding proofs. Indeed instead of just checking each line of a proof, one generally also wants to understand the proof "as a whole". This can be phenomenologically quite distinct from mere confidence that each line follows from the previous ones, as Tieszen (1992, pp. 58–59) emphasises. Though it is tempting to think about understanding in terms of the distinctive subjective subjective experience of grasping a proof, this characteristic sensation may not always be attainable – for instance it may be hard to gain the feeling of an immediate grasp of a proof taking longer than a page or so, or of a proof which relies on substantial previous results. An attractive alternative is to think of understanding as an ability, as advocated by Avigad (2011): for instance understanding a proof may mean that one can recreate it oneself, convey its ideas informally to another mathematician, use its ideas or techniques to prove similar results, generalize it, suggest how it could have been discovered, and so on. In many of these capacities that displays understanding of a proof, the ability to create or recognize valid proofs – proofs in which every inference is valid (as discussed above) – is key: recreating the proof means writing out a similar

valid proof of the same result, generalizing it means writing out a similar valid proof of a more general result, and so on. Arguably, to convey the ideas behind the proof to another mathematician means to use words, gestures, diagrams and so on to equip the recipient with the means to recreate for themselves a similar valid proof of the result in question. As Thurston (1994, pp. 31–32) emphasises, it can often be much easier to convey mathematical ideas by informal communication than by embedding them in and then excavating them from rigorous proofs.

A second point is that in practice in mathematics it is not demanded that every inference in a proof actually be valid. Total validity is intended when many philosophers speak of proof (as has been the attitude in this chapter), but in reality if an argument published in a mathematics journal contains a number of typos and minor logical errors, it may still be regarded as a perfectly acceptable proof. We could call the mathematicians' use of the word proof proof in the weak sense, to distinguish it from the philosophical use of the term. The key feature I think for proof in this weak sense is I think that the proportion of valid inferences is high, or very high, and that those inferences which are invalid are each fixable relatively easily. Thus for instance the classification of finite simple groups may well be a proof in this weak sense, even though it is so enormously long that it is practically certain that it contains errors. This is more or less the view taken by Aschbacher (2005). A proof of a result in this weak sense still establishes that its conclusion is a logical consequence of the basic principles used – since in principle one could convert it into a proof in the strict sense, in which case the result would be a logical consequence of the relevant basic principles (as seen in section I.7), and whether or not the conclusion is a logical consequence of the basic principles is independent of whether any such strict proof is actually written out.

Finally, a note on the epistemology of mathematics. In sections I.3 and I.4, an attempt was made to understand how rigour is judged in mathematics by thinking about how rigorous proof is learnt – by a gradual ascent up levels of greater and greater

compression, at each stage being able to tutor one's judgement of which inferences are valid by checking if they can be proved in greater detail. There is undoubtedly more that could be said about this process, but at any rate it is only part of a full epistemology of mathematics. Indeed it was argued in section I.7 that a valid proof shows its conclusion to be a logical consequence of the basic principles used; but this in itself does not imply that the conclusion is *true*. That would require further arguments, for instance arguments that the basic principles themselves are true. Thus the epistemology sketched here is one component of a full epistemology of proof, which would need to be supplemented by an epistemology of the basic principles themselves. Whether the basic principles generally used in mathematics – those of set theory – are true, and how we could know this, are exactly the kinds of epistemological questions that philosophy of mathematics has long wrestled with, and which some advocates of the shift to focusing on actual mathematics like to disparage, or describe as irrelevant to mathematics itself (for instance Rav 1999, Goethe and Friend 2010 and De Toffoli and Giardino 2016). It is of note that thinking about mathematical practice and the epistemology of proof leads us back naturally to exactly this standard epistemological question.

9 The choice of proof systems

The discussion of rigour in this chapter was centred around the case of set theory, formalized in the theory ZFC, but there is no requirement that rigorous mathematics be founded in this proof system. One could obtain a practice of high level rigorous mathematics founded on many other proof systems, such as the elementary theory of the category of sets (ETCS), or homotopy type theory; where in each case, once one has grasped how the basic principles of the proof system work, one can gradually elide more and more proof steps and move to proofs at levels of greater and greater compression, with proofs at each rung up in terms of compression always being able to be filled out

in greater detail to a previous level of compression already mastered. This is in many ways not a novel idea. Advocates of ETCS as a foundational system often claim that when one is familiar with the theory, and able to work using compressed informal proofs based on it in this manner, the reasoning that is available is essentially indistinguishable from that used in informal reasoning founded on ZFC; and one of the stated goals of the homotopy type theory (HoTT) book was to develop a new style of "informal type theory" (Univalent Foundations Program 2013, p. iv), or informal proof in HoTT, more compressed arguments founded in the formal theory in this kind of way. In each case, we can obtain by similar arguments as in section I.7 that more compressed arguments founded on a proof system are in principle formalizable in that proof system.

Thus we have a potential choice in mathematics about what proof system should be used to underlie our informal arguments – the choice of a foundation for mathematics. There are more options than those mentioned above, for instance buttressing ZFC with additional axioms, such as large cardinal axioms, or instead restricting to a potentially more secure proof system, such as a predicative system. In chapter VI we will consider what we should be looking for when choosing between such proof systems, and it will be argued that the key condition we should seek from a proof system is that it be sound – that when we can prove generalizations about some kind of mathematical structure in the theory, these generalizations do actually hold of all real examples of that kind of structure. First though, we turn to some case studies to test the account of rigour given in this chapter on.

Chapter II

Intuition and Permissible Actions

1 Alexander's lemma

Many authors disagree with the kind of analysis of rigour put forward in chapter I, where there is a link between rigour and formalizability. Two sets of objections have already been addressed: the idea that mathematicians should really be regarded as working in naive set theory (with unrestricted comprehension) in section I.6, and the worry about whether the process of filling in intermediate inferences in an argument might not terminate, in section I.7. A third kind of objection was mentioned at the end of section I.1: concern that a proof may involve reasoning that is somehow irreducibly high level, that resists being formalized, that requires radically new ideas perhaps to be introduced before it can be formalized. Although worries of this kind are often expressed (Rav 1999; 2007; Celluci 2009; Leitgeb 2009; Goethe and Friend 2010; Larvor 2012), the authors rarely give concrete examples of substantial real world mathematical arguments to demonstrate the phenomena they describe.

A prominent exception is the case of a famous argument from knot theory called Alexander's lemma (Alexander 1923). This received philosophical attention after Field's medallist Vaughan Jones recounted it in a philosophical piece (Jones 1998), describing it

as an easy, intuitive argument that would be very difficult to formalize. Following Jones, De Toffoli and Giardino make the argument the centrepiece of a case study (De Toffoli and Giardino 2016) in which they defend a rather different view of proof to that put forward in chapter I. De Toffoli and Giardino's claims have subsequently been taken up, with mild reservations, by Larvor (2019).[1]

Properly disentangling all the claims that have been made about this lemma takes some work. The situation is made complicated by the fact that the concept of tame knot – the key concept in the lemma – is standardly made precise in two very different ways: via polygonal knots, or via smooth knots.[2] Proving a theorem about polygonal knots allows one to deduce the corresponding theorem about smooth knots and vice versa, thanks to equivalence results relating the two concepts; but actually carrying out a proof in terms of polygonal knots can be a very different matter to carrying out a proof in terms of smooth knots, as turns out to be the case with Alexander's lemma. Alexander's original proof worked with polygonal knots, and Jones's version used smooth knots – and this apparently minor difference has huge consequences for the rigour, and formalizability, of the arguments.

In this chapter I consider De Toffoli and Giardino's analysis of proof (De Toffoli and Giardino 2016), which makes some claims that are in tension with the perspective from chapter I. They recount Alexander's lemma and use it as a case study to support their analysis of proof, but I argue that in fact their version of the argument mixes together features of both Alexander's original proof and Jones's retelling – and that this causes them to misunderstand what is going on in Alexander's proof at key points, leading them astray in some of their main conclusions about it, and about proof more generally. I argue that Alexander's original proof is actually a very good illustration of a rigorous

[1] Larvor (2012) also references Alexander's argument, but omits many of the strong claims made about it.

[2] Tame knots are a special case of the more general notion of knot, being those knots which are only "finitely knotted". Knots which are not tame are called wild. The theory of wild knots has not progressed nearly as far as the theory of tame knots, and they are not important in this thesis.

argument according to the standard described in chapter I.

That clarifies the status of Alexander's original proof, but not of Jones's version of the argument and the comments he makes about it – and leaves open the possibility that Jones's version of the argument could be used to support some of De Toffoli and Giardino and Larvor's claims, and attack the kind of account of proof put forward in chapter I. Chapter III considers Jones's version of the argument and defends against this possible attack, in the process drawing general lessons about what rigour requires of pictorial arguments.

Before proceeding, it is worth briefly stating Alexander's result itself. This result concerns tame knots, and there are different ways of making this concept precise – a fact which turns out to be crucially important for assessing the various versions of the argument, as mentioned above. For now though, a tame knot can just be thought of as a tangled, knotted loop of string in space – though one which is only "finitely" tangled and knotted (more precise definitions are found in section II.3). A key tool in knot theory is the ability to project a knot on a suitable plane, obtaining a knot diagram that indicates all its salient features (fig. II.1).

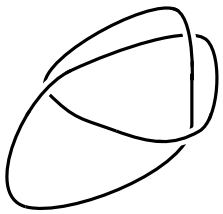


Figure II.1: A knot diagram

Alexander's lemma just states that every tame knot is equivalent to one with a diagram that only winds one way around an axis. Figure II.2 shows this to be true, as an example, of the diagram in fig. II.1.

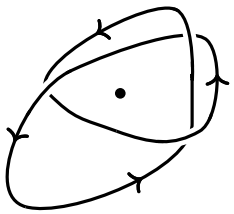


Figure II.2: A knot diagram winding around an axis

2 De Toffoli and Giardino's account of proof

Here I describe the key features of the account of proof that De Toffoli and Giardino put forward, and hope to support by considering Alexander's lemma. Some aspects of it I think are correct, but other parts I take issue with. Some of their claims about Alexander's lemma are echoed by Larvor (2019), as discussed here. Section II.3 then relates De Toffoli and Giardino's version of Alexander's argument.

One of De Toffoli and Giardino's main aims (following Larvor 2012, p. 716) is to

> challenge the model of formal logic as adequate to account for proof (De Toffoli and Giardino 2016, p. 27)

It is not totally clear what view they (and Larvor) are intending to counter here, however. Formal derivations are a model of proof, and how good or bad a model is will depend on what you are using it for. It is not clear whether anyone has claimed that formal derivations are the right model in all circumstances – this is not claimed by Azzouni (2013) or Burgess (2015), two recent defenders of a link between informal and formal derivations, and nor is it claimed by my account in chapter I. There are some ways in which formal derivations are obviously unlike the proofs mathematicians write: if all you know is the rules of natural deduction, there is no way you will be able to follow research level mathematics, no matter how smart you are. There is thus more to say about informal proof than just that it is modelled by formal derivations, and indeed chapter I sketched a simple account of how the standard of proof in much of mathematics

– rigour – works. Some of what De Toffoli and Giardino go on to say is compatible with that, and can be seen as useful additions to it for the particular case of low dimensional topology, or more widely. Some of their account is in contradiction with it though, and I will generally take issue with these parts. Some of their stronger claims in this direction are unsupported by Alexander's proof, as we will see in sections II.4 and II.5.

De Toffoli and Giardino are right to emphasise the collective aspects of mathematical practice (De Toffoli and Giardino 2016, pp. 28–29). They are also right to emphasise the harnessing of existing human cognitive capacities during mathematical reasoning (ibid., pp. 29–30). This would be included under what I vaguely termed "high level" reasoning in sections I.3 and I.4. However I would amend their discussion to emphasise that when it comes to rigorous mathematics, the important question is how the existing cognitive capacities are linked with judgements of provability: how does one learn reliably that certain natural ways of reasoning can be backed up (given the time and inclination) with proofs? They next discuss representations, in particular systems of notation, and I think they make important points here about the value of efficient, suggestive notation (ibid., pp. 30–32).

De Toffoli and Giardino's next topic of "permissible actions" is the main one about which I have reservations. The concept is drawn from Larvor (2012). To motivate their discussion of permissible actions, they appeal to a quote from Jones (1998):

> I remember being worried by Russell's paradox as a youngster, and am still worried by it, but I hope to demonstrate ... that it is not at all difficult to live with that worry while having complete confidence in one's mathematics (Jones 1998, p. 203; De Toffoli and Giardino 2016, p. 203)

They infer from this quote that confidence in mathematics is not based on "'logic' or foundations", and ask what the actual grounds for conviction are. It is worth saying a bit about this before moving on to discuss permissible actions. A basic point is that it is crucial to distinguish how one can gain *conviction* in mathematics from the question of

what the standard of *proof* is in mathematics. It is certainly true that rigorous proof is not the only way to gain conviction in mathematics: indeed this is the point of Jones's quote above, and he supports it with a number examples, such as his discussion of the huge number of applications of the Fourier transform, making the point that even if all our proofs of its properties turned out to be fallacious (or built on inconsistent assumptions) there must still be some sense in which this transform is true or valid (Jones 1998, pp. 203–204). In section I.2 we also saw a number of examples of inferences that may be convincing (and some of which were once accepted as valid), but would not be acceptable by the modern standard of rigour in mathematics. Though Jones is correct that conviction can be generated without a rigorous proof, that does not mean that that is the norm in mathematics, or that we should look elsewhere for the actual grounds for conviction; nor is this evidence either for or against any analysis of proof, whether based on logic or otherwise.

Now, onto the topic of permissible actions. De Toffoli and Giardino believe that mathematicians can gauge whether a proof is correct by seeing whether it consists entirely of these permissible actions, which are ways of reasoning that are accepted by the community of practitioners. As they put it,

> To become a practitioner means to learn to operate correctly on the representations, that is, to perform the appropriate actions. (De Toffoli and Giardino 2016, pp. 32–33)

They describe the proof as being addressed to this particular community of practitioners, a community which

> defines the 'permissible actions' on the representations. (ibid., p. 44)

They believe that when Alexander refers to "legitimate operations" he means these kinds of "permissible actions". They describe these as

> part of [the community's] mental model, [which] can be considered as reliable to gain new knowledge about the object of research. (ibid., p. 45)

They also introduce the term "local criteria of validity" in this connection, arguing that different areas of mathematics will have "different criteria of validity" (ibid., p. 49).

It is true that there are standards for what is acceptable in mathematical proof, and I would agree that there is no logic based criterion for this. An attempt to roughly describe how the standard of rigour in mathematics works was given in chapter I, and I did not put forward a criterion: one with no prior experience of evaluating mathematics could not read that account and hope to be able to judge the correctness of proofs. The distinctive feature of De Toffoli and Giardino's analysis is in seeing mathematicians as split into distinct communities, each with their own idiosyncratic ways of reasoning and their own local standards of correctness – standards which each individual community defines, without any further justification being supplied or called for. De Toffoli and Giardino write as though each community's ways of reasoning are automatically accurate about the community's chosen subject matter, because they form part of the mental model the community shares.

One obvious question this analysis ignores is where these communities come from. Practitioners of the various branches of mathematics have not been passing their wisdom down from one generation to the next since time immemorial. Most branches of modern mathematics have only existed in their present form since around 1900 or later, with the modern notion of mathematical rigour only stemming from around that time. It is not clear how the creation of new branches of mathematics and new mathematical communities would fit into De Toffoli and Giardino's account. They seem to be denying any general standard for what is acceptable in proof, which suggests that each community is free to set its standards as it likes on formation (though as De Toffoli and Giardino tell it, these standards appear to be fixed once they have been accepted by the community). Can any group of people studying mathematical subject matter call themselves a com-

munity of mathematicians, no matter how they do it? What if they extend the notion of proof to include numerical evidence, or conclusions reached in dreams?

This would evidently be problematic, and the reality is that the creation of new branches of mathematics is a routine part of the ordinary functioning of the subject. Indeed new branches of mathematics – studied by a particular "community" – are invented with some regularity. However one cannot just make up whatever kind of mathematics one likes, positing the existence of new kinds of objects, together with new basic principles stating how they behave: in rigorous mathematics, the birth of a new branch requires a demonstration of how its objects can be defined in terms of existing concepts, and how its basic principles can be demonstrated as consequences of these definitions. For instance associativity is in a sense an axiom of group theory, but we do not need to posit it as a new basic principle – it is just a property that (by definition) any group has.

There are occasionally what might look like exceptions to this, notably the axiom of universes sometimes used when working with categories (in particular in modern algebraic geometry, following Grothendieck). This is not an ad hoc assumption about categories however. One can be justified in appealing to it because (the feeling goes) it could perfectly well have been amongst the basic principles from the start. It can be precisely stated, can be motivated philosophically in a similar way to the other axioms, and is known to be independent of them. Having noted this special case, it can be put to one side.

It is true that in each branch there will be distinctive ways of reasoning, or "permissible actions". However De Toffoli and Giardino appear to suggest that these actions are reliable because the community accepts them – that since they are accepted they form part of the shared mental model of the practitioners, and thus are constitutive of the subject matter of the branch, so are automatically an accurate way to reason about that subject matter. In reality, in rigorous mathematics the opposite is the case. The permissible actions are not reliable because the community accepts them: the community

accepts them because they are reliable – because they can be seen and checked to be accurate ways of reasoning about the subject matter, according to the definitions given.

As well as being too permissive in its implications for what standards a community of mathematicians can set, the analysis in terms of permissible actions also does not properly reflect the pervasiveness and importance of novelty in mathematical arguments. De Toffoli and Giardino do accept the possibility that the practice of mathematics may evolve, for instance with material representations (symbols, notation, diagrams and so on) stemming from certain mental models, but then leading to insights which feed back in and modify the mental models themselves (De Toffoli and Giardino 2016, p. 30). But this is only a potential source of gradual change in the standard of proof that a community accepts, and it seems that at each point in time on this view there is still a fixed list "permissible actions" which states what kinds of inferences can be made, a list taught to each new practitioner as a student. If a novel kind of argument is made, not comprised of inferences on the list of permissible actions, then whether this argument is valid or not will (apparently) come down to whether the community can be persuaded to change their standards of proof to accept it.

In reality however mathematical reasoning is not nearly so constricted. Consider the introduction of probabilistic methods into combinatorics by Erdős, the application of linear algebra to group theory by Frobenius and others, and the development of homology by numerous mathematicians (including Alexander himself). If a brilliant mathematician develops a new way of reasoning about some object, then if that way of reasoning is correct, and can be seen to be correct, and justified in greater detail and precision if necessary, then it is a valid way of reasoning – even if the community had never even considered it before. Many breakthroughs in mathematics consist of exactly this. Even in more 'everyday' mathematics, papers will often contain new ways of arguing and new ideas, but on a smaller scale. The novelty we see in mathematics is possible precisely because there is a general standard for acceptable proof, one not

constituted by the methods each community of mathematicians currently happens to use.

De Toffoli and Giardino's account of permissible actions – and their attack on logic based approaches to proof – is supported by their view that visualization, imagination and intuition are crucial to the practice of proving in topology. They emphasise the importance of reasoning by envisioning and imagining transformations of topological representations (De Toffoli and Giardino 2016, pp. 44–46), and regard intuition as playing an ineliminable role at a key point in the proof (ibid., pp. 44). Larvor (2019) accepts their analysis in terms of visualizing transformations (pointing out a similarity with thinking about how one could twist a piece of rope around), and accepts their claim that "Alexander had no qualms about relying on [spatial intuition] in this proof", though noting that this is remarkable (ibid., p. 14). He does then row back somewhat however, pointing out that Alexander used a polygonal notion of knot and knot deformation in his argument, in contrast to De Toffoli and Giardino's account – an absolutely key point, as seen here and in chapter III.

Now the account of rigour from chapter I does not rule out uses of intuition and visualization; although the focus there was on the ability to prove inferences in greater detail, it was noted – in section I.4 – that it is also essential in mathematics that arguments can proceed at a high level, without one having a detailed proof of each inference in mind (either when writing, or when following the argument). Certainly some such high level inferences will involve visualization, imagination, and what might be called "intuition".

However for these kinds of judgements to be rigorous – as the notion is usually understood – this has to be intuition of a rather special kind. One cannot just be giving an untutored judgement of the plausibility of a claim: one has to be judging its provability. A classic example to illustrate this is the Jordan curve theorem, referenced in section I.2, which states roughly that every continuous injective closed curve in the

plane has an inside and an outside. This is intuitively about as obvious a statement as one can give, and to someone without experience of pathological functions it is probably hard to imagine what a counterexample could possibly look like. Nonetheless the proof is famously hard (if one works from the definitions, without tools like algebraic topology). Part of learning rigorous mathematics is learning to tell the difference between a statement like the Jordan curve theorem – which is obvious, but hard to prove – and a statement like the intermediate value theorem, which is obvious and whose proof is in fact straightforward. Of course intuitive judgements of plausibility are very important in mathematics: it is crucial that us humans are able to judge a statement like that of the Jordan curve theorem to be very likely, and thus set out to prove it. But when doing rigorous mathematics, there is a great difference between the kinds of judgements that would guide research in this way, and the kinds of judgements that are acceptable in a proof itself.

The rest of this chapter considers Alexander's argument to see whether it supports De Toffoli and Giardino's more controversial conclusions – in terms of each branch of mathematics having its own local standards of correctness, and topological proofs sometimes involving essential appeals to intuition – or whether it supports the kind of view of rigour seen in chapter I, some features of which were recalled here.

A first issue with their analysis – in which mathematics is split up into separate communities, with their own standards of proof and ways of reasoning – is that it is not even clear whether there was an established community of knot theorists at the time Alexander was writing (1923): this was before some of the major inaugural results of the field, such as Reidemeister's theorem – the seminal theorem which states that any equivalent tame knots have diagrams which can be related by a finite sequence of the three Reidemeister moves (Alexander and Briggs 1926; Reidemeister 1927).

It will be argued at any rate that Alexander's paper is not addressed to such a community, using ways of reasoning that only an initiate would understand. Instead the

entire argument is elementary, and straightforward to anyone with a basic knowledge of mathematics. He does not assume a background knowledge of knot theory, gesturing at a simple definition of tame knot as composed of a finite number of straight pieces (though he does not state this completely precisely) – here we see the concepts of the new field being defined in terms of existing concepts, as discussed above. Based on this definition, we can follow the argument, and see his inferences about tame knots to be accurate – not because we have been taught special kinds of reasoning used by knot theorists, but because we already have a grasp of how straight line segments in $\mathbb{R}^3$ behave. We can follow the argument involving the new concepts because of our grasp of the existing concepts, checking any inferences in greater detail as necessary. The argument Alexander gives is rigorous by the general standard enforced widely in mathematics, discussed in chapter I – it does not rely on some special standard of proof used by knot theorists.

It is true that one key concept in the argument – the "legitimate operations" – goes undefined, but it is clear from the context what this is intended to mean, as I discuss in section II.4. This is one critical point where De Toffoli and Giardino misinterpret Alexander, apparently taking him to be working with an intuitive notion of continuous (or smooth) transformation, without precise definition. This misinterpretation leads them to misunderstand the inference Alexander makes with the notion, which leads them in turn to overstate Alexander's reliance on intuition and visualization.

A second respect in which they misinterpret Alexander, also leading them to overstate his reliance on intuition and visualization, is in the structure of his argument: how his argument ensures that the process of knot modifications described terminates. Again they claim he is relying purely on intuition to justify this, and again their claim is erroneous (as a claim about Alexander's argument), as I discuss in section II.5.

First, I will briefly describe De Toffoli and Giardino's account of Alexander's argument, before looking at these two aspects in detail.

3 De Toffoli and Giardino's account of Alexander's argument

As discussed in section II.1, there are currently three versions of Alexander's argument in play: Alexander's original proof (Alexander 1923), a description by Field's medallist Vaughan Jones in a philosophical piece (Jones 1998), and the version of De Toffoli and Giardino in their own philosophical piece (De Toffoli and Giardino 2016). Alexander's and Jones' versions are importantly different, but De Toffoli and Giardino's version combines together aspects of both, and this is where the problems stem from.

Here I will limit myself to describing the key features of De Toffoli and Giardino's version. Later I will mention contrasting features of Alexander's original, but I will not relate his whole proof, as it is perfectly accessible in its original form – brief, simple and clearly written (Alexander 1923).

Now for some definitions. A *knot* is just defined to be a continuous injective map $S^1 \to \mathbb{R}^3$. Two knots are *equivalent* if they are related by an ambient isotopy, which is a continuous deformation of the first into the second which also deforms the ambient space continuously.

The kind of knots we are interested in are the tame knots, those which are only "finitely knotted". As discussed in section II.1, there are two different standard ways to define what a tame knot is – and this turns out to be important when unpicking what is going on in the different accounts of the lemma. The first is to use the notion of polygonal knot, a loop made of a finite number of straight line segments, intersecting only at the relevant endpoints. The second is to use that of smooth knot, a smooth non self intersecting map $S^1 \to \mathbb{R}^3$. It is the case that every polygonal knot is equivalent to a smooth knot and every smooth knot is equivalent to a polygonal knot; and we can thus define a *tame* knot to be one which is equivalent to a polygonal knot, or – equivalently – to a smooth knot.

Alexander works quite explicitly with polygonal knots, whereas Jones phrases his version of the argument for smooth knots.[3] Ultimately these give the same conclusion, since every polygonal knot is equivalent to a smooth knot and vice versa; but the arguments are (necessarily) quite different. De Toffoli and Giardino oscillate between regarding the knot they are discussing as polygonal and as smooth as they move through the argument, following Alexander in places and Jones in others, and this is how some crucial aspects of Alexander's argument are lost.[4] When discussing De Toffoli and Giardino's paper, Larvor notes that Alexander did work with polygonal knots, and wonders whether this might matter to their conclusions (Larvor 2019, p. 2728); he is right to wonder about this, though does not fully realise its importance.

Now onto the argument itself. De Toffoli and Giardino phrase this as showing that any knot is equivalent to a closed braid (see their paper for an account of braids, whose nature will not be important here). However they limit themselves to arguing for the result seen in section II.2, that any tame knot has a diagram in which there is an axis around which the knot always goes the same way – always clockwise or always anti-clockwise. This is Alexander's original lemma, which had no mention of braids – braids were only defined a few years later – though the fact that every tame knot has a representation as a closed braid is a quick corollary.

The relevant part of De Toffoli and Giardino's account of the argument (De Toffoli and Giardino 2016, p. 41) starts by taking a tame knot K with diagram $\mathcal{D}_K$, and taking this diagram $\mathcal{D}_K$ to be polygonal – thus implicitly assuming that K is polygonal (which they can do since any tame knot is equivalent to a polygonal knot). They take a small linear piece AB of $\mathcal{D}_K$ which does not contain more than one crossing, and choose a point C such that O lies in the triangle ABC. They replace AB in the diagram by the two segments AC and CB. This gives a precise description of the intended modification

[3] A full discussion of Jones's argument is left until chapter III, with a more detailed discussion of the definitions for smooth knots (involving a slightly different characterization of them) in section III.3.

[4] It is clear that De Toffoli and Giardino are aware though that smooth and polygonal knots are different, as seen for instance in their footnote 26 (De Toffoli and Giardino 2016, p. 41).

to the knot diagram, but leaves open the question of what the modification of the knot K is which leads to this change in $\mathcal{D}_K$. This is one of the key points where De Toffoli and Giardino depart from Alexander's proof.

To explain what modification of K gives rise to this change in $\mathcal{D}_K$, they appeal to a Jones's version of the argument. He is working with smooth knots, rather than polygonal knots, and phrases this key part by saying that one "throws it over one's shoulder", referring to the short stretch of knot being focused on (Jones 1998, p. 211). He illustrates this with a diagram like that of fig. II.3. De Toffoli and Giardino repeat Jones's phrase, saying that one throws the segment AB over one's shoulder (ibid., p. 211). They reference pictures and videos of how this manoeuvre could be carried out on a smooth knot, looking again like fig. II.3.

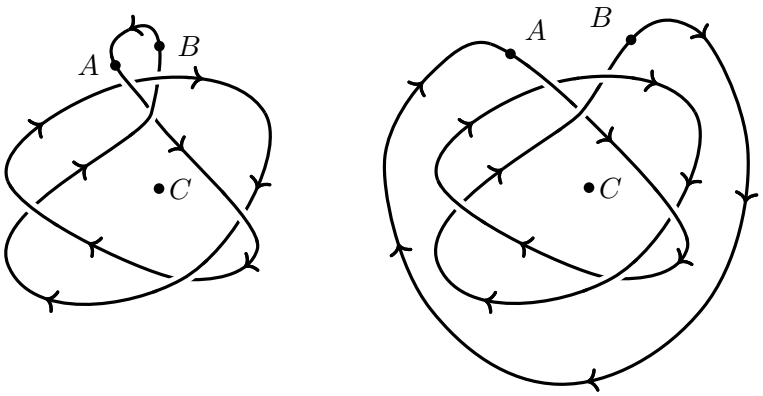


Figure II.3: The over the shoulder manoeuvre

Basing this part of their argument on Jones's version, their description and pictures of this manoeuvre involve smooth knots. This clashes badly with the context, as by assumption their knot is polygonal. They describe how

> Intuitively, the move consists in replacing a portion of the knot that goes in the opposite direction by throwing it in the other side of the point O so that it goes in the right direction. This has to be done carefully, without introducing new entanglements. (De Toffoli and Giardino 2016, pp. 41–42)

They do not describe why care is needed, how entanglements could be introduced, or how they could be avoided – and it is appears that again in this remark they are describing the smooth rather than polygonal case.

It appears they feel that in this description they are clarifying details left implicit by Alexander, quoting Alexander as saying

> the transformation of $\mathcal{D}_K$ *obviously* corresponds to an isotopic transformation of the space curve L (De Toffoli and Giardino 2016, pp. 42, emphasis De Toffoli and Giardino, notation changed by them)

Here they use L in place of K, as Alexander is discussing a system of linked knots (for which De Toffoli and Giardino are introducing this symbol L) rather than just a single knot.

Their final remark is that by repeating the process, one can eliminate every segment of the diagram which went in the wrong direction, and obtain the desired result.

There are two key respects here in which De Toffoli and Giardino unwittingly alter Alexander's argument. One, highlighted above, is in what modification is made to the knot K that leads to the described modification of the diagram $\mathcal{D}_K$. The second is in the structure of the argument, leading De Toffoli and Giardino to believe that intuition is required to deliver that the process of knot modifications terminates. These alterations are the source of De Toffoli and Giardino's bolder claims about the argument, which they use as grounds for their analysis of proof more generally. The first is discussed in section II.4, and the second in section II.5.

4 The "legitimate operations"

Firstly we have the nature of the knot modification Alexander uses: given (in De Toffoli and Giardino's notation) a knot K, we make some modification to it that corresponds to the transformation of the diagram $\mathcal{D}_K$ discussed in section II.3. When interpreting

this De Toffoli and Giardino drawing on Jones's version of the argument, despite Jones working with smooth rather than polygonal knots, meaning De Toffoli and Giardino's description and pictures make little sense with regard to the polygonal knot K they (and more importantly, Alexander) are working with.

De Toffoli and Giardino then misinterpret Alexander's phrase of modifying the knot using "legitimate operations". Working from their pictures and description – based entirely on Jones's version of the argument, and a recent video by Dalvit (2012) – they infer that Alexander is appealing to a shared practice amongst topologists of envisioning continuous transformations. They believe that this form of reasoning is not propositional and cannot be reduced to formal statements. They thus believe that Alexander's argument is not valid according to any general standard of validity that applies throughout mathematics, only being valid according to a special, local standard of validity (based on envisioning these kinds of continuous transformations) used in some areas of low dimensional topology. This is the main basis for their claim about mathematics being broken up into separate communities, each with their own standard of validity, that was discussed in section II.2. A second contributing factor to this claim is their altering the structure of Alexander's argument, discussed in section II.5.

To understand what Alexander actually means, it will help to make clearer the context of the relevant part of his argument. Firstly, Alexander is quite explicit that he is working with a polygonal notion of knot, assuming that a knot is composed of a finite number of straight line segments in $\mathbb{R}^3$ (Alexander 1923, p. 93). This is central to the way his proof works, as he moves through the finite number of straight line segments one by one, fixing any which in the diagram go the wrong way around the axis (section II.5 discusses the structure of his argument more closely).

When discussing Alexander's proof, I will base my notation on De Toffoli and Giardino's from section II.3, rather on Alexander's, but it is useful to supplement it. I will

write $[a_1, \ldots a_n]$ for the convex hull of $a_1, \ldots a_n$, defined to be

$$[a_1, \ldots a_n] = \Big\{ \sum_{i=1}^{n} \lambda_i a_i \mid 0 \leqslant \lambda_i \leqslant 1, \sum_{i=1}^{n} \lambda_i = 1 \Big\}.$$

Thus for instance $[a, b]$ is the line segment between point a and point b (for a, b distinct), and $[a, b, c]$ is the closed triangular region with a, b, c as its vertices (for a, b, c not collinear). If $a < b \in \mathbb{R}$ then this segment $[a, b]$ is the usual closed interval with endpoints a and b.

We have a polygonal knot K in $\mathbb{R}^3$, with projection $\mathcal{D}_K$ onto a plane P. Let $\pi : \mathbb{R}^3 \to P$ be the orthogonal projection, so $\mathcal{D}_K = \pi(K)$. We have picked a point O in P, and we are modifying $\mathcal{D}_K$ so that it only goes clockwise (say) around O. $[A, B]$ is a subsegment of $\mathcal{D}_K$ which goes anti-clockwise, and such that $\mathcal{D}_K$ has at most one crossing on $[A, B]$. We select a point C such that the point O lies in the interior of the triangle $[A, B, C]$. We seek to find a knot K' which is equivalent to K such that the diagram $\mathcal{D}_{K'}$ of K' is the same as $\mathcal{D}_K$, but with the two segments $[A, C]$, $[C, B]$ replacing the single segment $[A, B]$. This is the context of the quote from Alexander seen at the end of section II.3:

> The transformation of $\mathcal{D}_K$ obviously corresponds to an isotopic transformation of the space figure K. (Alexander 1923, p. 94)

(the notation here has been modified to fit with De Toffoli and Giardino's account[5]).

This is where De Toffoli and Giardino appeal to Jones's version of the argument, for the smooth case, using his phrase about throwing the knot over one's shoulder, with a diagram like that in fig. II.3. They also use stills from a a video by Dalvit (2012) made to illustrate the smooth version of the argument. As mentioned in section II.3 and at the start of this section, this makes little sense in the context Alexander is working. His knot is polygonal and no smooth isotopy can be applied to it (due to kinks in the knot where

[5]Alexander here is actually talking about a linked system of knots S, rather than a single knot K, but this is of no importance for us.

the different segments meet). Also, the kinds of continuous/smooth transformations that De Toffoli and Giardino describe and picture would not lead to a result with the required diagram – the same as that of K, but with the two segments $[A, C]$, $[C, B]$ replacing the single segment $[A, B]$. If one isotopied K into a smooth knot, the result would have a smooth diagram, not a polygonal diagram.

However if one puts aside Jones's version of the argument and instead focuses just on what Alexander is saying, it is clear what he means. We will suppose first that there is a single segment of K lying above $[A, B]$, so that there are $a, b \in K$ such that $[a, b] \subseteq K$ and $\pi([a, b]) = [A, B]$ (actually it appears to be an oversight by Alexander that this is not guaranteed at this point, as will be discussed later in this section; a slight rephrasing of the argument would guarantee this). If $\mathcal{D}_K$ has a crossing point on $[a, b]$, with $x \in [a, b]$ such that there is $y \notin [a, b]$ with $\pi(x) = \pi(y)$, then we can assume WLOG (by a rotation of space) that $x - y$ points vertically upwards. Thus the region vertically above the line segment $[a, b]$ is free from obstructions.

We are seeking a knot K' obtained by an isotopic transformation of K such that $\pi(K')$ is the same as $\pi(K) = \mathcal{D}_K$ but with the two segments $[A, C]$, $[C, B]$ replacing the single segment $[A, B]$. Thus K' must have the line segment $[a, b]$ replaced by some combination of line segments in $\mathbb{R}^3$ whose projection (under π) is $[A, C] \cup [C, B]$. So there must be a point c with $\pi(c) = C$, and a joined to c in K' by a sequence of line segments which project to $[A, C]$, and c joined to b in K' by a sequence of line segments which project to $[C, B]$. Does such a point c exist?

Obviously yes. As we are visualizing it, the region vertically above $[a, b]$ is free from obstructions, so if we take c to be enormously high up then the triangle $[a, b, c]$ will go almost straight up from the line segment $[a, b]$, and will not hit anywhere in K – in other words, with $[a, b, c] \cap K = [a, b]$. This is illustrated in fig. II.4. Thus we can take K' to be K but with $[a, b]$ replaced by $[a, c] \cup [c, b]$, which has the required projection, as seen in fig. II.5.

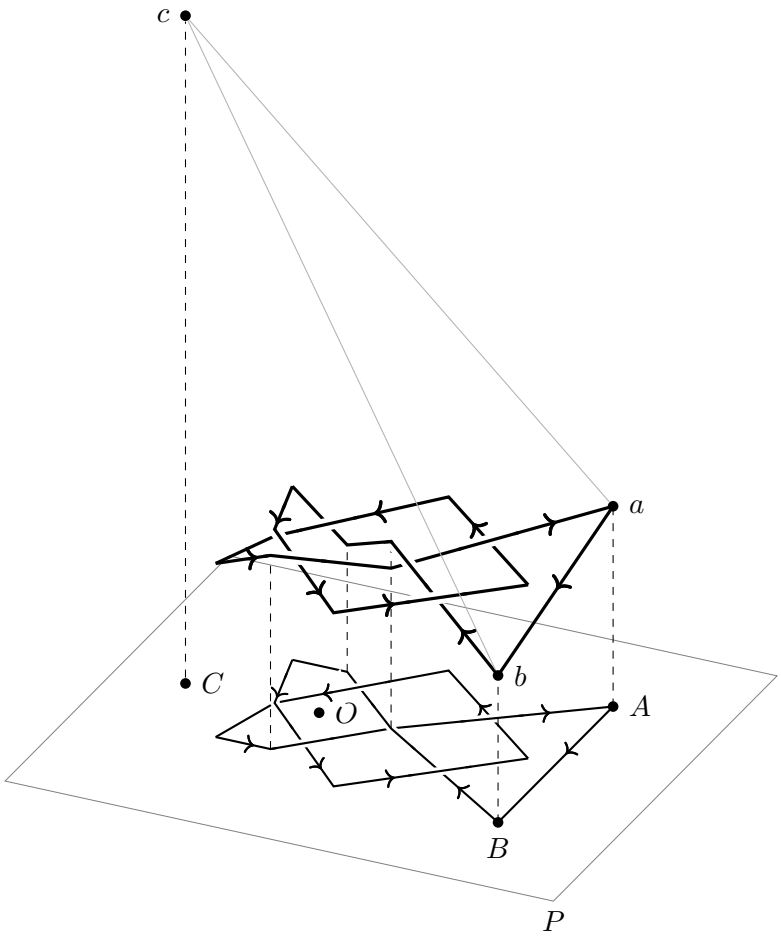


Figure II.4: Avoiding K with the triangle $[a, b, c]$

There is no question that this is what Alexander intends, rather than the vaguely specified continuous/smooth transformation De Toffoli and Giardino describe and picture. Perhaps they were attempting to make the proof more accessible to a lay audience, but in truth figs. II.4 and II.5 give a simpler and clearer picture than their account.

It is clear from the preceding Alexander has a notion of isotopy in mind on which if we have a knot K with a segment $[a, b]$ and a point c such that the triangle $[a, b, c] \cap K = [a, b]$, then K is isotopic to K' where K' is the same as K but with $[a, b]$ replaced by $[a, c] \cup [c, b]$. If Alexander had a notion of isotopy in mind on which this was not possible, his paper would be misleading at this key point. We don't need to know any more about his notion

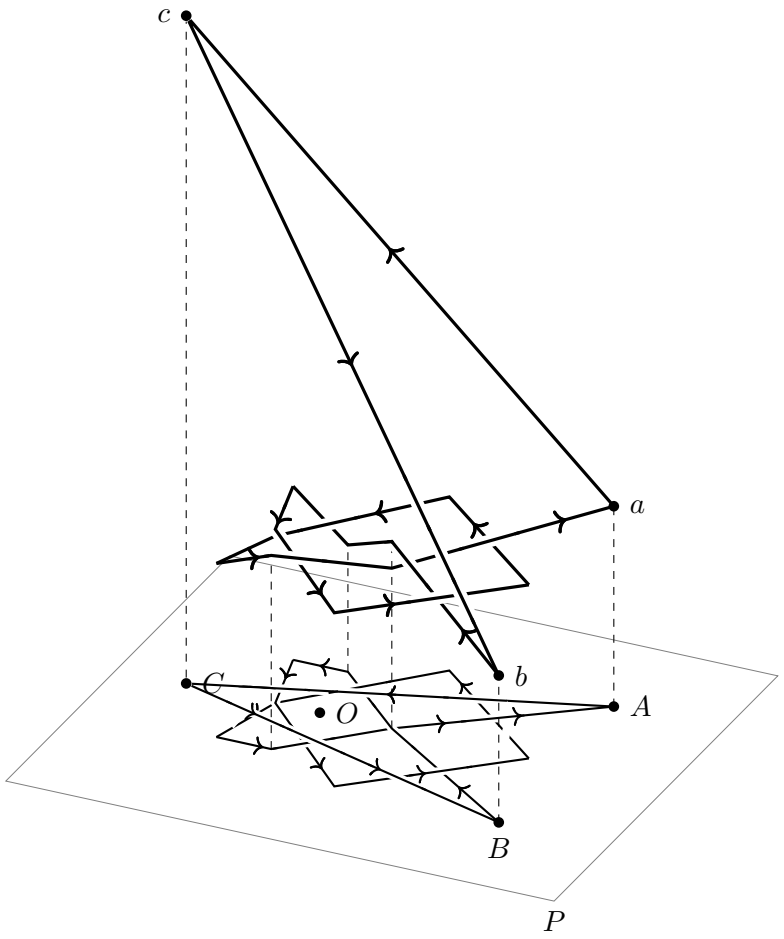


Figure II.5: The result of replacing $[a, b]$ with $[a, c] \cup [c, b]$

of isotopy than this to follow his argument, and this much we can infer from it.

It turns out that this is essentially exactly the standard notion of equivalence for polygonal knots. Alexander gives the definition in another paper:

> On any edge AB we may construct a triangle ABC, so drawn that neither the vertex C, the edge AC, the edge CB, nor the plane triangular region bounded by ABC has a point in common with the knot. We may then transform the knot by removing the edge AB and substituting in its place the edges AC and CB, along with the vertex C. We may also perform the reverse operation which consists in replacing a pair of consecutive edges AC and CB, together

> with their common vertex C by a single edge AB, provided neither the edge AB nor the plane triangular region bounded by ABC has a point in common with the knot. Each of the transformations here described will be called an elementary deformation. (Alexander and Briggs 1926, p. 563)

He defines two knots K_1 and K_2 to be of the same type if they can be related by a finite sequence of elementary deformations of the above kind. If this holds I will instead say that K_1 can be polygonally deformed into K_2. This is an equivalence relation.

The standard notion of equivalence for arbitrary knots (not just polygonal) is that of ambient isotopy. We define a knot here to be a continuous injective map $\phi : S^1 \to \mathbb{R}^3$. Then an ambient isotopy is a continuous map $H : \mathbb{R}^3 \times [0,1] \to \mathbb{R}^3$ such that $t \mapsto H(t,s)$ is a homeomorphism $\mathbb{R}^3 \to \mathbb{R}^3$ for all s and $H(t,0) = t$ for all t. If ϕ, ψ are knots, an ambient isotopy from ϕ to ψ is an ambient isotopy H such that $H(\phi(t),1) = \psi(t)$ for all t. We call ϕ, ψ ambient isotopic if an ambient isotopy from ϕ to ψ exists, and this is an equivalent relation on knots.

It is in fact the case that two polygonal knots are equivalent under polygonal deformation iff they are ambient isotopic. This is a basic fact of knot theory, in a sense more basic than the equivalence of smooth and piecewise linear notions of knot that De Toffoli and Giardino cite (De Toffoli and Giardino 2016, p. 41, footnote 26). One direction of this equivalence of equivalences is easy: that if K_1 and K_2 are polygonal knots such that K_1 can be polygonally deformed into K_2, then K_1 is ambient isotopic to K_2 (actually ambient isotopic via a piecewise linear ambient isotopy). This is proved for instance as one of the first propositions in Burde and Zieschang (2002, pp. 6–7, implication (3)⇒(2) of Proposition 1.10). Thus under either polygonal deformation or ambient isotopy, it is clear that replacing $[a,b]$ in K by $[a,c] \cup [c,b]$ gives an equivalent knot (the former by definition, the latter by a simple argument). Thus under either definition Alexander's proof is valid, and we do not need to know which one he intended to follow it.

I will shortly discuss how De Toffoli and Giardino's claims hold up in light of all these

points. Before that there are two things that should be remarked on. The first is the existence of a point c high enough up that $K \cap [a, b, c] = [a, b]$. This is a good example of the kind of high level reasoning discussed in sections I.3, I.4 and II.2. Someone trained in maths can "see" this to be true by visualizing the situation; but it is also clear how one would spell this out in greater detail. For $\lambda \geqslant 0$ let $c_\lambda = c + \lambda n$ where n is the normal to P pointing "upwards", i.e. in the direction of $x - y$ if K has a crossing point $x \in [a, b]$ with $\pi(y) = \pi(x)$, $y \notin [a, b]$, as discussed above (if there is no such x we can take n to be any non zero normal to P). Then the claim is that for λ sufficiently large, $K \cap [a, b, c_\lambda] = [a, b]$. We can split this up into multiple subclaims. Let $[d, a]$ be the edge of K preceding $[a, b]$, and $[b, e]$ the edge following $[a, b]$. Let $[p, q]$ be the edge containing y if there is such a y, otherwise we can take $[p, q] = \varnothing$. Then we have that

$$K \setminus ((d, a] \cup [a, b] \cup [b, e) \cup (p, q))$$

is compact with

$$\pi(K \setminus ((d, a] \cup [a, b] \cup [b, e) \cup (p, q))) \cap [A, B] = \varnothing,$$

and we need to argue that:

- For λ sufficiently large, $[a, b, c_\lambda] \cap [d, a] = \{a\}$
- For λ sufficiently large, $[a, b, c_\lambda] \cap [b, e] = \{b\}$
- For λ sufficiently large, $[a, b, c_\lambda] \cap [p, q] = \varnothing$
- For λ sufficiently large, $[a, b, c_\lambda] \cap (K \setminus ((d, a] \cup [a, b] \cup [b, e) \cup (p, q))) = \varnothing$.

Each of these can indeed be proved in greater detail if necessary. It appears that this comes to a few pages, if written out comprehensively. Of course one does not have to write this out to see Alexander's proof to be valid; but it is important for rigour that

it be possible to argue the inference in greater detail if called for, and that it is not an irreducibly high level intuition. As discussed in section I.3 and section II.2, I think there is an important question for the epistemology of mathematical proof here: how and in what circumstances can one gain the ability to reliably judge high level inferences like this to be provable in greater detail?

The second point is that in the above, I introduced the assumption that there is a single segment of K lying above $[A, B]$ – that there are $a, b \in K$ such that $[a, b] \subseteq K$ and $\pi([a, b]) = [A, B]$. If one was careless when visualizing the situation one might well assume that $[A, B]$ would have to have such a line segment $[a, b]$ lying above it, but in fact this need not be the case: all we can guarantee is that there is a finite sequence $[a_1, b_1], \ldots [a_n, b_n]$ of line segments contained in K with $[A, B] = \bigcup_{i=1}^{n} \pi([a_i, b_i])$. Each of these line segments $[a_i, b_i]$ must lie above the line segment $[A, B]$, but they can have different vertical components to their gradients. In this case the argument proceeds much the same way as above, but the point c has to be picked high enough that for each i, the triangle $[a_i, b_i, c]$ only intersects K in $[a_i, b_i]$. The unnecessary complication this creates appears to be a simple oversight by Alexander. When he talks about "P mov[ing] along certain segments of the broken line" (Alexander 1923, p. 94) he could just as easily talk instead about P moving along the projection of certain segments of the knot above. This would not affect the rest of his proof at all, and in this case each segment $[A, B]$ like the one we considered above would have a single segment $[a, b]$ of K above it.

Now to De Toffoli and Giardino's claims about this part of the argument. First, they claim that the reasoning is not propositional reasoning, nor formal reasoning, and is not based on formal reasoning, nor can it be reduced to formal statements (De Toffoli and Giardino 2016, pp. 43–44, 48–49). It is not entirely clear what they mean by this. Alexander's written proof consists entirely of words and symbols, and contains no pictures – in what sense is it "not propositional"? One can reason from a first proposition

to a second in many different ways, including via one's spatiotemporal faculties. Perhaps when they say propositional reasoning, they mean reasoning in terms of strict logical rules; and of course Alexander's argument is not literally a formal argument – nor are most published proofs. This is not a significant point though, and I doubt anyone has ever claimed the opposite. Although Alexander's proof is not formal, as discussed in chapter I and above it is important for rigour that its inferences be provable in greater detail if requested; and this is indeed the case, as sketched for one key inference above. If one keeps repeating this process, asking for greater detail/more precision in every inference, and then for greater detail/more precision in each of those more detailed inferences in turn, one will eventually reach a formal derivation. This is line with the briefly sketched argument in section I.7 that all rigorous proofs are formalizable, as a consequence of the norm of rigour. I am not claiming any epistemic benefits to this here however, just noting that it can be done.

With regard to the claimed importance of non-propositional reasoning, it is also worth noting that the crucial clarifications of Alexander's argument given above were propositional – the correct intended knot modification, the existence of the point c high enough above the knot that $K \cap [a, b, c] = [a, b]$, and so on. These propositions can be illustrated visually, but if one had to limit oneself to the propositions or the illustrations in writing out the argument, I think the propositions would be the part to keep.

Although Alexander's proof would require a normal mathematician to do some visualising to follow it, De Toffoli and Giardino do not quite grasp the nature of the visualization involved. They describe Alexander's proof as based on the manipulation of concrete spatio-temporal objects (ibid., p. 44), which is inaccurate as Alexander's proof is based on knots being a finite union of straight line segments, which are not concrete and have zero width (of course in some crude sense one would could trace back a grasp of how straight lines behave to familiarity with concrete objects, but in this sense almost all mathematical reasoning would be based on the concrete and the claim is uninteresting).

They repeatedly refer to Alexander's argument as involving smooth or continuous transformations (De Toffoli and Giardino 2016, pp. 41, 43, 44, 45, 46). As discussed above, Alexander intends a polygonal deformation of the knot; referring to this as "continuous" is misleading in its excess generality, and referring to it as "smooth" is incorrect. This polygonal deformation requires a much more straightforward visualization than the continuous/smooth ones they indicate in their various diagrams (ibid., pp. 42). Their remarks about being careful not to introduce new entanglements while transforming the knot might be pertinent to Jones's version, but are not relevant to Alexander's actual proof with its simple polygonal transformation (ibid., p. 42).

This all leads them to overestimate the role played by visualization in the proof, which is much simpler and more easily backed up by detailed arguments then they describe. This completely undermines their claim that Alexander's proof relies on a special "local" standard of validity used by topologists, in terms of envisioning continuous transformations (ibid., pp. 43–46, 48–49). In fact Alexander's proof is perfectly rigorous by the usual standards in mathematics (with a mild imperfection in the point noted above that he should guarantee a single segment of K lying above $[A, B]$, but does not, unnecessarily complicating things slightly).

De Toffoli and Giardino's claims here are really only suited to Jones's version of the argument – they are right that Jones's version seems to have fairly irreducible appeals to intuition and spatiotemporal reasoning, and that it would be very difficult to prove in greater detail or formalize. A full discussion of Jones's version of the argument, and whether it can be used to support these comments on proof, is found in chapter III.

5 Termination of the process

There a second respect in which De Toffoli and Giardino misrepresent Alexander's argument which leads them to overstate its reliance on visualization and intuition. The

proof describes a sequence of modifications to a knot, and it is essential to the proof that this sequence eventually terminates, in a knot with a diagram of the required form (only going the right way around an axis in the plane); if it does not terminate, the lemma fails. Here De Toffoli and Giardino claim that

> it is left to our intuition to prove that ... it is not an infinite process. Alexander does not really [give] us any other justification: this reasoning plays an epistemic role. (ibid., p. 44)

However as was the case in section II.4, their conclusion rests on a confusion. In this case, they miss out key steps in Alexander's reasoning, which ensure the termination of the process. They are wrong to think that in their version of the argument "intuition" could guarantee the termination of the process – in the argument as they have stated it, there is no guarantee that the process will terminate.

I will start by discussing the problem with De Toffoli and Giardino's version of the argument. They choose a small straight portion $[A, B]$ of the diagram, which goes the wrong way around O and contains at most one crossing, and they correct this one segment – bending it to go the other way around O. They then move onto another small straight portion of the diagram which goes the wrong way and contains at most one crossing, and do the same. Since the diagram has only finitely many crossings, one might hope that this process would always terminate. The problem is that when bending a segment to go the right way, one may introduce *extra* crossings to the diagram, and one may in fact introduce extra crossings to the troublesome parts of the diagram – that go the wrong way around O. Thus one could potentially keep on going forever, bending more and more segments of the diagram to go the right way, but constantly adding to the workload as one goes by increasing the number of troublesome crossings. The diagram would get more and more complicated, with smaller and smaller segments being bent the right way each time. The lemma would fail.

De Toffoli and Giardino show some awareness of this problem in the above quote, but

are wrong to think that it can be brushed aside by "intuition" (De Toffoli and Giardino 2016, p. 44). In the process as they describe it, they have left the above possibility wide open. It would not be difficult to describe a sequence of knot modifications that fits their description but never terminates: take two troublesome sections S_1 and S_2 on opposite sides of O, and first correct a section of S_1 containing a crossing while simultaneously adding at least one troublesome crossing to S_2, then correct a section of S_2 containing a crossing while simultaneously adding at least one troublesome crossing to S_1, and so on. Of course one could use one's "intuition" to see that this could be avoided – that one could give a more careful description of the process that ruled out this possibility. But that is not to use intuition to see their argument is valid: it would be to use intuition to rewrite their argument to make it valid. Their comment that one has to carry out the over the shoulder manoeuvre "carefully" to avoid introducing new entanglements (ibid., p. 42) does not help, since the procedure being described is one that has to work without human oversight or intelligence (it has to work just as well for a knot diagram with 10^{1000} crossings as with 10).

In fact, the problem is easily avoided, as seen in Alexander's actual proof. The key difference between his proof and the version described by De Toffoli and Giardino is in its logical structure – exactly the kind of feature that a perspective focused overmuch on visualization and intuition is likely to miss. Alexander's proof is not an induction, which is the attempted structure of De Toffoli and Giardino's; it is a *double* induction, with the part of the argument described by De Toffoli and Giardino being the inner induction.

Indeed, Alexander's proof first considers the set of segments of $\mathcal{D}_K$ which bend the wrong way around O (in his notation, he considers the set of segments of S_π which bend the wrong way around L). I will call this set T here. His argument deals with each element σ of T in turn, by breaking each such σ up into finitely many subsegments $\sigma_1, \ldots \sigma_n$ on each of which there is at most one crossing. The point is that one when one corrects the subsegment σ_i one does not add crossings to σ – though one may add

crossings to other elements of T. To make this completely clear, we can phrase the argument as follows. I will not be entirely formal here, sufficing to make clear this double induction structure.

Proposition. *Suppose K is a polygonal knot and σ a line segment contained in $\mathcal{D}_K$ which goes around O the wrong way. Suppose σ_i is a subsegment of σ such that $\mathcal{D}_K$ has at most one crossing on σ. Then K is equivalent to a polygonal knot L with the same diagram as K, except with the subsegment $\sigma_i = [A, B]$ replaced by two segments $[A, C]$ and $[C, B]$ with C a point such that $O \in [A, B, C]$.*

Proof. This is the part of the argument discussed in section II.4, and the part that appears in De Toffoli and Giardino's account (in somewhat altered form, as discussed in section II.3 and section II.4). □

Proposition. *Suppose K is a polygonal knot and σ a line segment contained in $\mathcal{D}_K$ which goes around O the wrong way. Then K is equivalent to a polygonal knot L which has the same diagram as K outside of σ, and such that L's diagram goes the right way around O on the part it replaces σ with.*

Proof. We break σ up into subsegments $\sigma_1, \ldots \sigma_n$ such that each σ_i has at most one crossing. Then by repeatedly applying the previous proposition (this is the inner induction) to each σ_i in turn, we obtain the result. Here we use the fact that if $\sigma_i = [A, B]$ and C is a point such that $O \in [A, B, C]$, then $([A, C] \cup [C, B]) \cap \sigma = \{A, B\}$, so that replacing σ_i with $[A, C] \cap [C, B]$ does not add any crossings to any σ_j for $j > i$. □

Proposition. *Suppose K is a polygonal knot. Then K is equivalent to a polygonal knot L with a diagram which only goes around O the right way.*

Proof. This is by induction on the size of the set of segments of $\mathcal{D}_K$ which go around O the wrong way, with the previous proposition providing the induction hypothesis (and the base case trivial). □

Thus Alexander's argument here is perfectly rigorous – by the normal standards – as stated. The apparent flaw De Toffoli and Giardino discuss, the possibility that the process need not terminate – which they look to intuition to solve – is a flaw their version inherits from Jones's, and has no root in Alexander's original argument.

In summary, the account of rigour put forward in chapter I is unthreatened by Alexander's proof. On the contrary, Alexander's proof is a good illustration of it. All of De Toffoli and Giardino's stronger claims about Alexander's argument rest on two alterations: concerning the nature of the knot deformation Alexander intends, and the structure of his argument. With these points cleared up, their claims about his argument are seen to have no basis. The grounds for Larvor's statement (echoing De Toffoli and Giardino) that "Alexander had no qualms about relying on [spatial intuition] in this proof" are also removed.

This resolution of De Toffoli and Giardino's claims about Alexander's argument destroys the basis for their more general claims about mathematics being split into different communities, each with their own standard of validity, claims which were critiqued in section II.2. The remaining question is whether Jones's argument can be used to support some of De Toffoli and Giardino's comments and attack the kind of account of rigour put forward in chapter I – a question considered in the next chapter.

Chapter III

Rigour, Pictures and Knot Theory

This chapter has two linked goals. Firstly, to clarify the status of Jones's version of Alexander's lemma; and secondly, to determine what rigour in general requires of pictorial arguments (a loose end from chapter I).

Jones describes the argument he puts forward as being very simple and intuitive, but very difficult to formalize. De Toffoli and Giardino (2016) make these same claims about Alexander's lemma, echoed in turn (with a little hesitancy) by Larvor (2019); as seen in chapter II, De Toffoli and Giardino believe they are analyzing Alexander's original proof, but in fact they misrepresent key aspects of Alexander's argument by mixing it with Jones's. The possibility remains that De Toffoli and Giardino's, and Larvor's comments are accurate when applied to Jones's version however. If correct – if Jones's argument is rigorous, simple, intuitive and very difficult to formalize – this could pose a challenge to the view of rigour put forward in chapter I.

However this chapter contends that the reason the argument is so hard to formalize is because it falls a long way short of being rigorous, by the normal standard. To show this, the chapter uses typical features of what is normally understood by rigour in

mathematics, rather than directly relying on the account from chapter I; in the process, though, the close connection between the account from chapter I and these usual features of rigour is illustrated.

What rigour requires of pictorial arguments in particular is a question left unanswered by chapter I. Many authors believe that pictorial reasoning may prove a problem for the kind of view of rigour put forward there, arguing reasoning of this kind may provide good examples of valid yet unformalizable arguments (Leitgeb 2009; Goethe and Friend 2010; De Toffoli and Giardino 2015; 2016; Larvor 2019). This chapter considers the case of pictorial arguments, and argues – based on the example of Jones's argument – that a crucial requirement for them to be rigorous is that we be able to state which features of the pictures are actually used in the argument, and which are merely accidental features of how the pictures happen to be drawn. This is used in section III.8 to defend the standard view of proof against the general objection that pictorial arguments may be valid yet unformalizable, and also to provide an amendment to Larvor's analysis of what is required of pictorial arguments for them to be rigorous (Larvor 2019).

To make Jones's argument rigorous much mathematical work needs to be done. Here I argue for various modifications that need to be made, and state some definitions and propositions that play a role in this. Many of the mathematical details are omitted, and are gathered in Appendix A. Most of the proofs are routine, but some contain ideas which are worth noting, and are discussed here. Propositions have the same numbering here and in section A.1.

1 Rigorous use of pictures I

We start by discussing what rigour requires of pictorial arguments – which might involve a sequence of pictures with inferences from one to the next, or might just have a single picture with inferences connecting it the to surrounding prose. What do we need of these

kinds of arguments for them to be rigorous? The claim I want to make is that we need to be able to state what features of the pictures are accidental – just features of they happen to be drawn – and which features actually play a role in the argument.

For a first example of an argument which satisfies this condition, we will consider the converse to the mutilated chessboard problem. The original mutilated chessboard problem is a famous brain teaser in which two opposite corner squares are removed from a chessboard, and it is asked whether it is possible to tile the remainder of the chessboard with 31 dominos, each covering two adjacent squares. Actually, this turns out to be impossible. When dominos are placed on a chessboard, each covers exactly one black square and one white square, so that placing a series of dominos on a board will always cover an equal number of black and white squares. But two opposite corner squares of a chessboard are the same colour, so once they are removed the board has a different number of black than white squares; thus it cannot be tiled by dominos. By the same argument whenever two squares of the same colour are removed from a chessboard, the result cannot be tiled by dominos.

The converse mutilated chessboard problem considers the opposite situation: what happens when one removes two squares of different colours? Then this impossibility argument fails. A mathematician Ralph Gomory gave a beautiful proof that in this case, tiling the result is always possible (Honsberger 1978, pp. 66–67). The proof essentially consists of the picture in fig. III.1.

As seen in fig. III.1, the squares of a chessboard can be put in a Hamiltonian cycle through adjacent squares, and thus with successors in the cycle having opposite colours. Then if squares A and B of opposite colour are removed there are an even number of squares between them in this order, which can then be tiled with dominos. The path this follows is shown in fig. III.1. Corners are no problem since a domino can be placed either vertically or horizontally at them as required.

Not only is this a very clever and pleasing argument, it is also perfectly rigorous. It

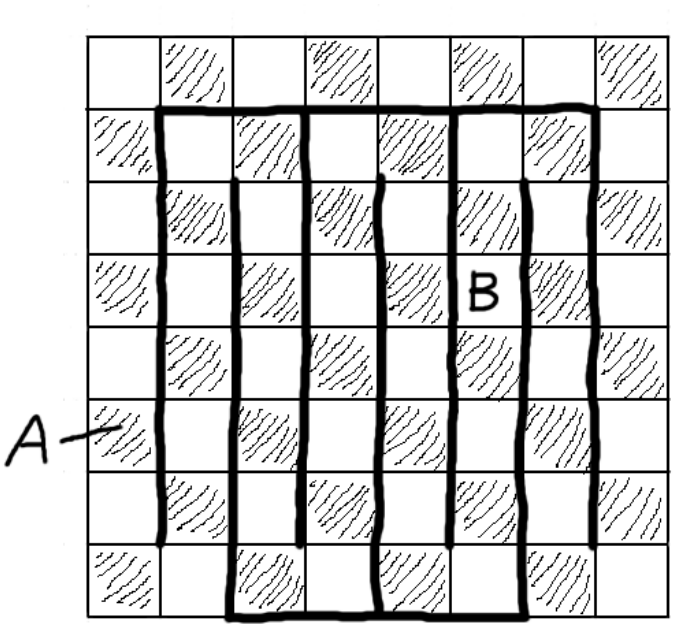


Figure III.1: Removing squares of different colours from a chessboard

is completely clear what the picture shows: a Hamiltonian cycle through the chessboard squares, with successive squares adjacent. It is also clear that some aspects of the picture are mathematically irrelevant. For instance the shading on the black squares is not identical, with some having more dense shading in places than others – but this one can very safely ignore. This is almost too obvious to be worth pointing out, except that this clarity is not always present: with some pictures, we may not be clear on which aspects are relevant to the argument, and which are just accidents of presentation. As we will see, this is the case with Jones's argument – one reason for its lack of rigour.

Since we can clearly state what the content of the diagram fig. III.1 is – and what inference we make with it – this argument fits the paradigm of rigour described in chapter I. Suppose one had an undergraduate student who is thoroughly immersed in the detailed, careful manner of proof of analysis, and hesitates about the above discussion, wanting to know how to carry it through in a more careful style. It is not difficult to

make things more precise. If S is a set of chessboard squares, with $A \subseteq S^2$ the set of pairs of adjacent squares, then from fig. III.1 we can see that there is a bijection $f : \mathbb{Z}_{64} \to S$ where for each $n \in \mathbb{Z}_{64}$ we have $(f(n), f(n+1)) \in A$. Here $\mathbb{Z}_{64}$ denotes the quotient group $\frac{\mathbb{Z}}{64\mathbb{Z}}$ of integers modulo 64. This bijection from a cyclic group to the chessboard squares is one way of being more precise about the intended Hamiltonian cycle. Other parts of the proof can be made more precise in similar ways if need be, with the picture still always serving as a visual aid.[1] Full formalization, including formally defining the above bijection f, might well be very tedious but probably also fairly routine.

The view that rigorous use of pictures requires that one can state which features of them are argument relevant can also be supported by considering the example of Euclidean geometry. Mumma (2010) discusses how though Euclid's diagrammatic reasoning was originally regarded as the paragon of rigour, during the 19$^{\text{th}}$ and 20$^{\text{th}}$ centuries it came to be seen as less rigorous than sentential formalizations of geometry. A driver of this was a lack of clarity over what kinds of inferences could be licensed by diagrams. The problem was made acute by examples where paradoxical conclusions can be reached using apparently valid diagrammatic reasoning (ibid., pp. 261–262). To dispel these worries one can distinguish between the *exact* properties of a diagram and the *coexact* properties: examples of the former are equalities or inequalities between angles and lengths, and examples of the latter are containment relations between regions of the diagram. The latter are stable under small perturbations of the diagram whereas the former may not be. Any diagram one draws will have necessarily have exact properties – one line segment being longer than another for instance – but on the analysis of Mumma (following Manders 2008) these are just features of how the diagram hap-

[1] Some might object that one should *not* spell out the details of a visual proof in precise terms like this, and should urge the student to meditate on the argument, let it slosh around in their mind until the right level of certainty is reached. My opinion is that if there is ever doubt about a higher level, quick proof, then – if one has the time – it can be very helpful to see how it can be made more precise and detailed, and thus seeing *why* the argument is valid. By doing so one's judgement for what higher level arguments are valid, and how to make valid high level arguments, can be continually sharpened. This point is emphasised by Tao (2009).

pens to be drawn, and are not used directly in inferences. One can only directly infer coexact properties of a diagram. This is part of why general conclusions can be drawn from particular diagrams, and is also a sensible condition since if one reproduces very slightly wrong then its exact properties may be lost. If one does want to deduce an exact property, one reasons propositionally using other exact properties as hypotheses, and perhaps coexact properties of a relevant diagram (Mumma 2010, pp. 262–267). The situation is further complicated by the presence of a construction stage during a diagrammatic argument, of drawing new lines, circles or points on a diagram. This leads to conditions on which coexact properties can be inferred (ibid., pp. 267–277).

The above is a very brief summary, but the basic point is the same as that of the chessboard example. For reasoning using Euclidean diagrams to be rigorous, we have to be able to state what kinds of inferences we can or can't draw from the diagrams: which features of the diagrams can be relied on, and which are merely accidental. Being able to spell this out is a major advance in our understanding of Euclid's reasoning, and confirms its rigour.

Jones's argument will be used to illustrate what it looks like when this requirement fails. Indeed in section III.5 it will be argued that we do not know what features of his diagrams are actually supposed to be playing a role in the argument; as a result, we are not sure what inferences are being made, or what key terms mean. This is a major bar to considering the argument as rigorous. This analysis of what rigour requires of pictorial arguments is returned to in section III.8, where it is used to amend Larvor's account of this topic. In between, detailed discussion of this and other aspects of Jones's argument takes place.

2 Jones's argument

This section recounts Jones's argument and his comments on it, and discusses those comments of De Toffoli and Giardino, and Larvor, which could be taken to use it as a basis. It also indicates those parts of the argument whose rigour will be discussed in the remainder of the chapter.

Unlike Alexander, Jones takes a knot to be a smooth closed curve in $\mathbb{R}^3$ (which should be non self intersecting, though he does not state this). Thus he is working with smooth knots, a special case of tame knots (as usually understood, and as defined in section II.3). He takes the notion of equivalence for such knots to be that of one being smoothly deformable into the other (Jones 1998, p. 209). Exactly what this means will be discussed in section III.3.

Jones's argument is very simple. Given a knot one wishes to make wind a certain way around an axis, one follows the knot round until one finds a stretch that is going the wrong way. Then

> One isolates a short stretch going the wrong way and "throws it over one's shoulder" until it is on the other side of the knot, going around correctly. (ibid., p. 211)

This is illustrated in fig. III.2.

Jones says that

> The only thing that could go wrong in the above procedure is that, in trying to throw a bit of string over one's shoulders, one may meet a crossing. This is easily handled. Since we are proceeding one short stretch at a time around the knot, simply isolate that crossing and, if it happens to prevent our throwing over our shoulders, throw it the other way. When we have arrived back at the beginning point, we [stop]. (ibid., p. 211)

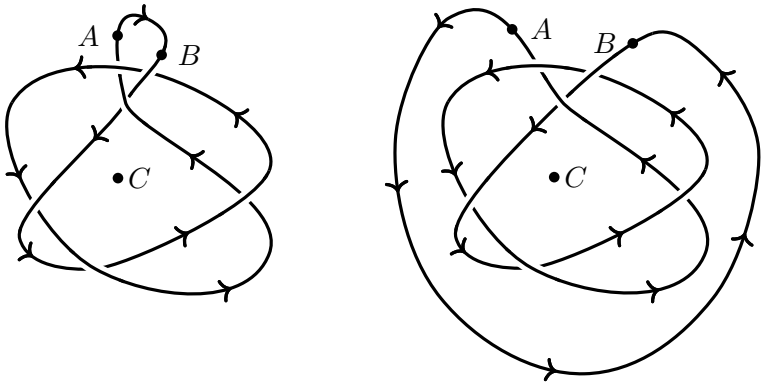


Figure III.2: The over the shoulder manoeuvre

That ends Jones's argument.

Jones contrasts this simple argument with an argument of Von Neumann's from functional analysis. He states that the theorem from functional analysis is difficult, requiring significant mathematical background to even understand, and many hours of work to fully appreciate the proof. By contrast this argument from knot theory is "easy": "the result and its background could be explained quiet rapidly to a clever high school student" (Jones 1998, p. 212). He continues:

> a careful analysis of these proofs reveals that the proof of [the knot theoretic result], if properly formalized, would be much longer than that of [the result from functional analysis]. One would have to be precise about the kinds of continuous deformations that are allowed, and constructing the functions required for the "throwing over the shoulder" trick would be a nightmare. (ibid., p. 212)

He states that the reason this knot theoretic result is "so easy" is because "we are able to bring to bear our full intuition about three-dimensional space on the problem" (ibid., p. 212).

Many of De Toffoli and Giardino's comments on Alexander's lemma echo Jones's comments on his version of the argument. They claim that the argument involves reasoning that "cannot be reduced to formal statements without completely altering the

proof" (De Toffoli and Giardino 2016, p. 48), emphasising the importance of visualization in the argument, and the lack of any further justification beyond the visual for some parts of it: they state that one key inference is "is left to our intuition to prove", that "Alexander does not really gives [sic] us any other justification: this reasoning plays an epistemic role" (ibid., p. 44). They mention both Jones and Alexander's versions of the argument and appear to believe that these conclusions apply to both. However as I argued in chapter II, they are mistaken in this. Alexander's original proof (Alexander 1923) is very different to Jones's retelling. The original is perfectly rigorous by normal standards: it has no big leaps of reasoning, and no essential appeals to intuition. Alexander works in the polygonal setting rather than the smooth setting which Jones uses, and this makes rigour much easier to achieve. De Toffoli and Giardino are misled in their analysis by Jones's retelling of the argument, and these comments of theirs only make sense when applied to Jones's version.

The same goes for much of Larvor's discussion of the argument. In Larvor (2012, p. 727) he accepts Jones's description of Alexander's proof, stating that the central move is the "over the shoulder" manoeuvre and that giving a physical demonstration of this – for instance with a piece of string, or on a chalk board – is the core inferential act of the proof. Then in Larvor (2019) he appears to largely assent to De Toffoli and Giardino's account of the argument, again repeating that the "over the shoulder" manoeuvre is the core of the proof, visualized as carried out on an imaginary loop of rope (ibid., pp. 13–14). As mentioned in section II.2, he accepts their claim that "Alexander had no qualms about relying on [spatial intuition] in this proof", though noting that this is remarkable (ibid., p. 14); and he does later row back somewhat, pointing out that Alexander used a polygonal notion of knot and knot deformation in his argument, in contrast to De Toffoli and Giardino's account. As seen here and in chapter II, this is an absolutely key distinction.

One central aspect of De Toffoli and Giardino's account cannot be supported by

Jones's version of the argument, however: their contention that each branch of mathematics has its own local criteria of validity, rather than there being a single dominant standard of rigour that (pure) branches collectively follow. Jones's version of the argument was not published in a mathematical journal, but is a proof sketch in a piece of philosophical musings. The focus of his article is on how conviction is generated in mathematics, and many of his examples are clearly cases where one can have conviction even in the absence of anything that would normally be called a proof – for instance in his discussion of the vast number of applications of the Fourier transform, which (he says) surely secure it from any attacks of logical inconsistency or contradiction (Jones 1998, pp. 203–204). His description of Alexander's lemma is informal, perhaps deliberately so to make it accessible to his philosophical audience, and he doesn't claim that the argument as he relates it would be acceptable to a mathematics journal – nor does he use the word "rigour" or its cognates anywhere in describing his argument. Additionally, Alexander's original proof of the lemma, which was published in a journal (Alexander 1923), is completely rigorous by normal standards, as seen in chapter II.

If Jones's argument is rigorous, and these comments represent a properly rigorous understanding of it, then that tells against the view of rigour from chapter I. Although that view did allow for high level inferences – without a detailed proof in mind – the key feature was that one's judgement of the validity of such inferences be tutored by an ability to prove them in more detail, allowing one to correct and clarify one's high level intuitions whenever necessary. In contrast these authors advocate an understanding of Jones's argument which is irreducibly high level, based solely on some intuitive grasp of how loops in space can be bent and shaped.[2] When they discuss the immense difficulty of formalizing the argument, what is really at stake is the difficulty of proving its assertions in any greater detail – to even get started on the process of formalization: to get a sense of how difficult this would be, just have a think for yourself about how you would give

[2]Jones, for instance, discusses how the argument could be relayed to and understood by a bright high school student, who clearly would have no more sophisticated understanding of the argument than this.

a more detailed, precise justification of any part of it. Thus apparently the only way to understand the argument is in terms of the kinds of irreducibly high level visualizations and intuitions these authors advocate.

One of the aims of the rest of this chapter is to defuse this threat: it will be argued that in fact Jones's argument badly fails to be rigorous by the normal standard. Of course it would be circular to use the account of rigour from chapter I to judge this, and instead standard features of the concept of rigour in mathematics will be appealed to. In the process, the closeness of the account of rigour from chapter I to what is normally meant by rigour in mathematics will be illustrated.

First, section III.3 discusses rigorous definitions of some central concepts of knot theory. After that detailed consideration of Jones's argument starts. Section III.4 discusses a major potential ambiguity in the form the sequence of knot modifications takes, arguing that Jones rather oversimplifies this in crucial respects, and that this leads De Toffoli and Giardino to misunderstand the structure of the argument. Rigour requires greater clarity here, and a rather more complex argument. Then section III.5 examines Jones's term "short stretches", and argues that it is ill defined in various important respects, and we need a more precise understanding of the term for the argument to go through. Here Jones's pictures are not much help, also being ambiguous in important ways – in that we do not know which features of the pictures are actually intended to be playing a role in the argument, and which are merely features of how the pictures happen to be drawn. Making Jones's argument rigorous requires a more precise definition of what the "short stretches" are, and thus also a clearer understanding of which features of the pictures are actually argument relevant. Having done this work, Alexander's lemma follow from a few clearly stated intermediate results. Of these the "over the shoulder manoeuvre" is key, and section III.6 argues by comparison with other results that rigour requires a proof of this, rather than just an appeal to intuition.

The details of all the definitions and proofs seen in these sections to be necessary to

make Jones's argument rigorous are given in Appendix A, and putting them all together gives us an argument which is rigorous by usual standards, and also rigorous by the standard described in chapter I: it is clear that every inference in it could be justified in more detail if required, and ultimately that the proof could be made formal without any great insight or ingenuity required, though perhaps requiring quite a bit of time due to its length. Section III.7 sums up the analysis of Jones's argument, and addresses the comments of Jones, De Toffoli and Giardino and Larvor. Finally section III.8 returns to the discussion from section III.1, arguing based on consideration of Jones's pictures in section III.5 that a key requirement for a pictorial argument to be rigorous is that we be able to state which features of the pictures are actually relevant to the argument, and which are merely there accidentally. This is used as the basis of a proposed amendment to Larvor's account of what rigour requires of pictorial arguments (Larvor 2019).

3 Definitions

Before examining the rigour of Jones's argument, we need to look at the concepts involved. As discussed in section I.2, for a mathematical definition to be rigorous it must be clear that the definition could be made formal in such a way that all uses of the concept would be valid. This is a totally standard feature of rigour that I assume Jones, De Toffoli and Giardino, and Larvor would accept.[3] At any rate there are various well known ways to give simple definitions for the basic concepts of knot theory, that I am sure these authors are familiar with. One can motivate the need for such rigorous definitions by considering natural questions about knots that would be left open by an intuitive characterization.[4]

[3]It is true that De Toffoli and Giardino speak rather vaguely about the "legitimate operations" used in the argument to deform the knot (De Toffoli and Giardino 2016, pp. 41, 44, 45, 48, 49), but they also mention that knots are considered up to ambient isotopy (ibid., p. 33), a concept they give a completely precise account of in another paper (De Toffoli and Giardino 2014, p. 831).

[4]For instance, can a knot be nowhere differentiable? Can it have positive area (like an Osgood curve)? Can it be infinitely knotted?

The option Jones takes is to regard knots as smooth. There are various ways of spelling this out, and for a rigorous argument we should settle on one. We will take smooth knots to be certain T-periodic curves for some $T > 0$. A map ϕ with domain $\mathbb{R}$ is T-periodic if we have $\phi(t) = \phi(t + T)$ for all $t \in \mathbb{R}$. It follows immediately that $\phi(t) = \phi(t + kT)$ for all $k \in \mathbb{Z}$. We will fix some $T > 0$ to serve as the periods of our curves. For $t, t' \in \mathbb{R}$, we write $t \equiv t'$ if there is $k \in \mathbb{Z}$ with $t' - t = Tk$, i.e. if $t' - t \in T\mathbb{Z}$. This is an equivalence relation on $\mathbb{R}$.

Definition III.1. A **smooth knot** is a smooth map $\gamma : \mathbb{R} \to \mathbb{R}^3$ with period T such that for all t, $\gamma'(t) \neq 0$, and such that γ is injective up to $\equiv$–equivalence – i.e. for any $t, t' \in \mathbb{R}$ we have $\gamma(t) = \gamma(t')$ iff $t \equiv t'$.

Taking knots to be T-periodic is simpler than taking them to have domain $[0, T]$, since then we would have to make sure that the derivatives at 0 matched up with the derivatives at T. We require that $\gamma'(t) \neq 0$ for all t since otherwise the knot could have kinks – $\gamma(t)$ could slow down and then stop as t approaches t_0 from below, and then start off again in a different direction as t increases beyond t_0. If B is a subset of $\mathbb{R}^3$, we denote the set of smooth T-periodic maps from $\mathbb{R}$ to $\mathbb{R}^3$ with image in B by $C^\infty_T(\mathbb{R}, B)$.

We also need a notion of equivalence for smooth knots. The one Jones gestures at is that of smooth isotopy (this is equivalent to other standard notions of equivalence)

Definition III.2. Let β and γ be smooth knots in $\mathbb{R}^3$. A **smooth isotopy** from β to γ is a smooth map $H : [0, 1] \times \mathbb{R} \to \mathbb{R}^3$ such that if we set $H_s : t \mapsto H(s, t)$ then H_s is a smooth knot for all $s \in [0, 1]$, and $H_0 = \beta$ and $H_1 = \gamma$. We write $\beta \sim \gamma$ if a smooth isotopy from β to γ exists.

Another key definition is that of *regular* diagram, a well behaved diagram (definition A.1.3). One can argue from the definition that such diagrams only have finitely many crossing points (theorem A.3.10), so are potentially suitable for being pictured as literal human diagrams.

One benefit of giving these kinds of rigorous definitions is that it allows one to bring more sophisticated mathematical ideas to bear on the problem. For instance, one can define a norm on the space of smooth knots which corresponds to the intuitive idea of when two smooth knots are close, and can be useful for proving facts about them.[5]

For Alexander's lemma we need a point in the plane around which our knot diagrams will wind. We will just take this to be the point 0 in the complex plane. We then define a quantity $D_\gamma(t)$ that indicates the direction that the diagram of a smooth knot γ goes around 0 (for the details, see the discussion following proposition A.1.5). This quantity is strictly positive or strictly negative when γ is winding anticlockwise or clockwise respectively around 0 at t (and is 0 when$\gamma_{\mathbb{C}}$ is going towards or away from 0 at t, or is stationary). It can be checked that this coincides with other ways of making rigorous the concept of which direction $\gamma_{\mathbb{C}}$ is going around 0.

Now we can give a precise statement of the result we aim to prove.

Alexander's lemma. *Let γ be a smooth knot. Then there is a smooth knot β which is smoothly isotopic to γ such that β has regular projection avoiding 0, and we have $D_\beta(t) > 0$ for all t.*

4 Structure of the argument

Now to discuss Jones's argument itself. The first basic thing to clarify about this argument is its structure. At a first glance, one might think that the argument consisted of dividing the knot diagram up into short stretches on which it goes the wrong way and has at most one crossing, then going through these short stretches and correcting each in turn with the over the shoulder manoeuvre. This is apparently how De Toffoli and Giardino interpret it (De Toffoli and Giardino 2016, pp. 41–42).

[5]The definition follows definition A.1.2. Under this norm, smooth isotopy classes of knots are open (theorem A.3.8). This captures the intuitive idea that if you perturb a smooth knot, you get another smooth knot equivalent to it. Mathematically, it means that to show that every knot is smoothly isotopic to a knot of some class L, it suffices to show that L is dense in the set of smooth knots.

However on this reading the argument fails: there is nothing to stop us from adding new crossings elsewhere as we go, potentially adding to the future workload rather than decreasing it. One could easily describe a sequence of knot modifications along these lines that did in fact go on forever. Just take a knot diagram with two stretches U_1 and U_2 on which it goes the wrong way around the axis, with U_1 and U_2 on opposite sides of the axis and each having multiple crossings. One starts by fixing a stretch of U_1 with only one crossing, and in doing so adds more crossings to the unfixed part of U_2; then one fixes a stretch of U_2 with only one crossing, and in doing so adds more crossings to the unfixed part of U_1, and keeps going in this fashion.

A naive reaction might be that this would be a very stupid way to proceed, but that is to misunderstand the nature of the argument required. We are not describing a procedure for a human to follow, so that given a knot they can – with some intelligence – obtain a diagram which only winds one way around some axis. I am emphasising this because De Toffoli and Giardino seem to think that by intuition one can make sure that the process terminates:

> it is left to our intuition to prove that this [sequence of knot modifications] is not an infinite process. (ibid., p. 44)

But the procedure we create is only as smart as we make it, and has no intuition of its own; and intuition cannot deliver termination of the process De Toffoli and Giardino describe, since as seen above termination is not guaranteed.

The key idea for a process of knot modifications which does terminate is very simple, and completely explicit in Alexander's original proof (Alexander 1923, p. 94). The idea is to structure the proof as a nested induction. First we pick out stretches $U_1, U_2, \ldots U_n$ on which the knot diagram goes the wrong way. Then we divide U_1 into shorter stretches $V_1, \ldots V_m$ on which it has only one crossing. We go through the V_i one by one, using the over the shoulder manoeuvre to fix V_i *without adding crossings to any unfixed* V_j, but perhaps adding crossings to U_i for $i > 1$. In this way we can fix all of U_1, and then move

onto U_2. We now divide U_2 into shorter stretches on which it has only one crossing, and fix each of these shorter stretches without adding crossings to unfixed parts of U_2, but perhaps adding crossings to U_i for $i > 2$. By continuing in this way we will fix each U_i, and thus obtain a knot diagram of the required form. This is illustrated in fig. III.3.

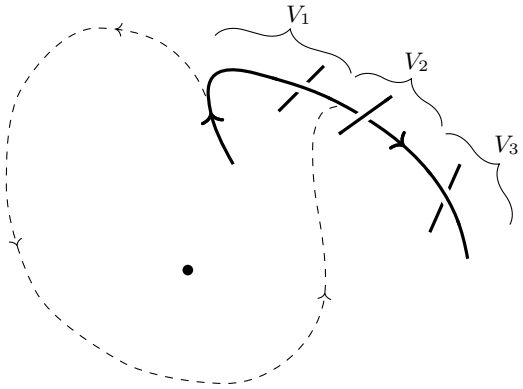


Figure III.3: The over the shoulder manoeuvre, avoiding unfixed short sections

If this argument seems harder to follow than what Jones describes, I sympathise. It is significantly more complicated. It has a crucially different logical structure, a nested induction unlike the failed simple induction argument sketched above. In fact I think this nested induction is what Jones intends: he talks first about finding a "stretch that is going the wrong way", and then divides this further into "short stretches" on which it has only one crossing. However his account is very unclear on this point. He makes no mention of the need to avoid adding crossings to the stretch one is fixing when throwing bits of it one by one over the shoulder. It may be that this is the kind of detail that he feels is so obvious that it does not need mentioning, but leaving it out is likely to lead to confusion – as it did for De Toffoli and Giardino, who missed the nested induction completely in their account, making their appeal to intuition instead to supposedly guarantee termination of the process. Thus to make the argument rigorous, we should be explicit about what is going on here.

It is also rather misleading, I think, to skip over these considerations. Jones is

trying to make a point about how straightforward an argument can be while being very difficult to formalize: but to make this point he leaves out a key aspect of it (and one which significantly complicates it).

5 “Short stretches”?

In our new sketch of the argument, we first divide the knot diagram into stretches $U_1, \dots U_n$ on which it bends the wrong way, and go through fixing these one at a time: we break the stretch U_i we are working on into shorter stretches on which it has at most one crossing, and one by one throw these shorter stretches over the shoulder to fix them, taking care not to add any crossings to unfixed parts of U_i, as illustrated in fig. III.3.

This is only a sketch though. Forgetting for the moment whether we will want to justify the inferences in this argument in greater detail, right now we do not even know what these inferences are. There are two main sources of uncertainty. One is the nature of these stretches U_i on which the knot diagram bends the wrong way – what properties do they have? The second is the over the shoulder manoeuvre – when we do this to the knot, what are we doing? What properties does the result have to have?

These two questions are connected. For the argument to work, we have to be able to cover troublesome parts of the knot diagram with finitely many of these U_i, which have certain nice properties. We will then be working through these U_i one at a time, repeatedly using the over the shoulder manoeuvre on it to correct it and bend it the right way around the axis. But when we are doing this to U_i, we may well affect the diagram on U_j for some $j > i$: these sets $U_1, \dots U_n$ may well overlap. Thus whatever nice properties the U_i have, the over the shoulder manoeuvre needs to be done on U_i in such a way that it doesn't spoil the nice properties of U_j for $j > i$.

Thus identifying the nice properties that the U_i must have is a somewhat delicate task: we need strong enough nice properties that we can carry out the over the shoulder

manoeuvre without difficulty, but we also need to avoid making them so strong that U_j's nice properties might be spoilt by carrying out the over the shoulder manoeuvre elsewhere on the knot. We also need to pick the properties of the U_i in such a way that troublesome parts of the knot diagram can be covered by finitely many of them.

One could be forgiven if one missed this point when reading Jones's exposition, as De Toffoli and Giardino (2016) do. Jones ignores the possibility that these stretches might overlap (Jones 1998, p. 211); indeed, he only considers the much easier special case in which there is a single U_i with a single crossing on it.

Carrying out this balancing act – making the nice properties of the U_i strong enough, but not too strong – is one of the most delicate parts of the proof. This kind of task is exactly what we would normally expect from someone carrying out a rigorous argument: as discussed in section I.2, a central requirement of rigour is that it be possible to give rigorous definitions of the concepts one uses – definitions that are clear, and could be made formal if requested. Here, however, Jones's prose and pictures are of minimal help.

The problem is essentially that mentioned in section III.1: that we are not sure which features of Jones's pictures are relevant to the argument, and which are just artefacts of how they happen to be drawn. As will be seen, there are a great range of candidate properties that the U_i should have, all compatible with his pictures, and he gives us no way to distil out those which are actually intended to be part of his argument.

One description Jones does give of the U_i is as being a "stretch", but that does not get us far. It is not a mathematical term and has little in the way of obvious mathematical meaning. We are talking about subsets of $\mathbb{R}$, and presumably when we describe them as "stretches" we are implying they are connected, and thus are intervals. Beyond this not much is clear though.

It seems like taking the U_i to be open may be a good idea, to allow us to make a compactness argument to obtain our finite list $U_1, \ldots U_n$. We can introduce a notation

for troublesome parts of the curve: if $A \subseteq \mathbb{R}$, we let

$$D_\gamma^{\leqslant 0}(A) = \{t \in A \mid D_\gamma(t) \leqslant 0\}$$

be the set of "troublesome" points in A (we define other similar notations such as $D_\gamma^{>0}(A)$ in the obvious way). Our aim in Alexander's lemma is to obtain a knot β smoothly isotopic to our initial knot γ with $D_\beta^{\leqslant 0}(\mathbb{R}) = \varnothing$, or equivalently with $D_\beta^{\leqslant 0}([0,T]) = \varnothing$. Since D_γ is a continuous function, $D_\gamma^{\leqslant 0}(A)$ is always a closed subset of A and so $D_\gamma^{\leqslant 0}([0,T])$ is compact. Thus if we can cover $D_\gamma^{\leqslant 0}([0,T])$ with a family of open intervals $(U_i)_{i\in I}$, then we can cover it with finitely many of the U_i. Note that even this basic property – that the U_i be open – is not obvious from Jones's diagrams; to the contrary, on his sketch it looks as though the endpoints A and B are included in the troublesome stretch to be corrected (ibid., p. 211).

A simplistic interpretation of the requirement that the diagram of γ bends the wrong way on each of these stretches U might be that we have $D_\gamma(t) \leqslant 0$ for all $t \in U$. But we cannot necessarily cover $D_\gamma^{\leqslant 0}([0,T])$ with such intervals, since we may have a point $t \in D_\gamma^{\leqslant 0}([0,T])$ with $D_\gamma(t) = 0$ and $D_\gamma(s) > 0$ for s approaching t from below (say).

Instead, it makes sense to require (roughly) that such a U contains an interval on which $D_\gamma \leqslant 0$, an interval which might be all of U. Thus γ might not "bend the wrong way" on U, but instead "has at most one backwards bend" on U. We will not need to really worry about the portion of U on which γ bends the right way.

The next property we need is that $\gamma|_U$ doesn't wind too far around the axis. Take a look at fig. III.3. If the V_i kept winding clockwise all the way round the knot, it could well be impossible to throw V_1 over the shoulder without hitting some V_i with $i > 1$. What we want is some control on the argument of $\gamma_{\mathbb{C}}$ on U, or more specifically on the part of U on which $\gamma_{\mathbb{C}}$ goes clockwise around the axis, i.e. $D_\gamma^{\leqslant 0}(U)$. We do this by requiring that $\gamma|_{D_\gamma^{\leqslant 0}(U)}$ bends at most half way clockwise around the axis, or more

formally that

$$\sup\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(U)\} \leqslant \inf\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(U)\} + \pi,$$

where Arg_γ is a smooth argument function for γ (see the discussion following proposition A.1.5 for the details of this).

There are other properties that might seem relevant. For instance, when picking the U_i, do we want them to be small enough that $\gamma_{\mathbb{C}}$ is roughly straight on them? We know that later when doing the over the shoulder manoeuvre we're going to want to bend part of U_i round as in fig. III.3 without hitting the unfixed portion of U_i, and this may be easier if U_i is roughly straight (imagine trying to carry out the modification in fig. III.3 if the V_j were enormously wiggly, oscillating in close to the axis point and then out again). What properties the U_i need to have will depend on how the rest of the argument is going to be carried through, which makes the lack of precision in this regard a more serious problem.

As it happens a "rough straightness" property like this will not be needed however. The properties listed above turn out to be sufficient. Putting them together, we obtain the following definition.

Definition III.3. Let I be a nonempty compact interval in $\mathbb{R}$. Let γ be a smooth knot which projects onto $\mathbb{C}$ avoiding 0. We say that γ has **at most one backwards bend** on I if $\sup(I) < \inf(I) + \frac{T}{2}$, and $D_\gamma^{\leqslant 0}(I)$ is an interval such that

$$\sup\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(U)\} \leqslant \inf\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(U)\} + \pi.$$

Then if J is any nonempty bounded interval we say that γ has **at most one backwards bend** on J if it has at most one backwards bend on $\overline{J}$.

This is independent of the choice of Arg_γ. We require $\sup(I) < \inf(I) + \frac{T}{2}$ so that

two such sections can only intersect at one end modulo T-equivalence, which simplifies things a little. The definition is carried out for compact intervals first because to carry out the proofs it turns out that having at most one backwards bend on $\overline{U}$, rather than just on U, is useful.

In a similar way we then need to work out what we are aiming to achieve when we are doing the over the shoulder manoeuvre – what properties should the bent segment have? One can go through a process like the above of considering potentially useful properties, and it turns out that the following is what's needed.

Definition III.4. Let γ be a smooth knot which has regular projection avoiding 0. Let I be a compact interval with $\sup(I) - \inf(I) < \frac{T}{2}$. Say that a smooth knot β is a **bending forwards of γ on** I if it has regular projection avoiding 0, is smoothly isotopic to γ and:

(i) $D_\beta^{>0}(I)$ is an interval

(ii) If $t \notin \mathrm{Int}(I) + T\mathbb{Z}$ then $\beta(t) = \gamma(t)$

(iii) $D_\beta(t) \geqslant D_\gamma(t)$ for all t

As discussed above, forming these definitions and making that they interact in the appropriate way is the most delicate part of making Jones's argument rigorous. It is here that the weakness of Jones's pictures is particularly clear. It is true that in a sense Jones's pictures do (roughly) display the properties listed above (Jones 1998, p. 211). For instance, the stretch between A and B he picks out is an interval and does bend at most half way around the axis, satisfying the key conditions of definition III.3. Then when this stretch is bent forwards, the part of the modified interval where the diagram goes the right way around the axis is itself an interval, so the condition "$D_\beta^{>0}(I)$" of definition III.4 is satisfied, and the condition "$D_\beta(t) \geqslant D_\gamma(t)$" is also visibly satisfied.

So his diagrams are compatible with the properties listed above: however that is little help, since they are also compatible with a vast array of other properties. In his diagram

the stretch from A to B isn't too wiggly – does this matter? The condition that we pick stretches which are roughly straight was considered above and actually, in the end, is not important to the argument. Jones's diagram is no help there. Similarly, the stretch from A to B has roughly constant radial distance from the axis – is this important? No, it turns out. Does it matter that the stretch from A to B has no parametric inflection points (points where it goes from bending "to the right" to bending "to the left", or vice versa)? Again, no, this is unimportant. In Jones's diagram the stretch from A to B is bounded on either side by stretches on which the knot diagram goes the right way (clockwise). Should we require this? Actually, no, we can't – since we are requiring that the stretches we pick bend at most half way around the knot axis (so may need to overlap with other stretches where the knot diagram goes the wrong way). One could keep listing such properties for ever.

The basic problem is that we are not sure which features of Jones's diagrams are a part of the argument, and which are accidental. The ability to make this clear was highlighted as a key requirement for an argument involving pictures to be rigorous in section III.1, and will be discussed further in section III.8. When Jones talks about picking the short stretches, and shorter stretches, it simply is not clear from his pictures what he means – what properties these stretches have. The same goes for the properties of the over the shoulder manoeuvre. As discussed above clarifying these is one of the subtlest parts of rigorizing the argument, since we need properties for the U_i strong enough that we can carry out the over the shoulder manoeuvre without being so strong that the over the shoulder manoeuvre spoils them.

So one major lack of rigour in the argument is that Jones's pictures are insufficient to pin down the key concepts – so that we do not even know what claims the argument involves. Once these concepts are properly defined we still have to make the argument though. Some basic properties of these concepts are straightforward, for instance the following.

Proposition III.5. *Suppose β is a bending forwards of γ on I and α is a bending forwards of β on J, such that there is $t \in I \cap J$ with $D_\beta(t) > 0$ and such that* $\sup(I \cup J) - \inf(I \cup J) < \frac{T}{2}$. *Then α is a bending forwards of γ on $I \cup J$.*

Then a crucial property we need, as emphasised above, is that when one "bends forwards" according to the above definition one does not mangle sections on which the knot has at most one backwards bend.

Proposition III.6 (Bending forwards does not interfere with other sections)**.** *Suppose γ is a smooth knot, and I is a compact interval such that γ has at most one backwards bend on I. Suppose that β is a bending forwards of γ on J such that that $D_\beta^{>0}(J) \nsubseteq \mathrm{Int}(D_\gamma^{\leqslant 0}(I)) + T\mathbb{Z}$. Then β has at most one backwards bend on I.*

The proofs of these propositions are routine. It is important that they can be proved though. We are certainly guided by intuition and visualization when forming these concepts and reasoning about them, but to definitively establish their properties we want a proof. It is not enough that one just cannot think up a counterexample – it could be that our imagination was limited in some way (as has often turned out to be the case in the history of mathematics). In this case our intuition that these facts are true leads naturally and straightforwardly to rigorous proofs. If somehow such a statement had turned out to be very hard to prove, that would have cast doubt on our initial intuition, and would indeed make providing a proof more urgent.

There is one more complication before we can proceed to the main part of the argument. This is that a smooth knot γ may not have a diagram which can be covered by open intervals on which it has at most one backwards bend, since it may wiggle between going clockwise and anticlockwise round the axis infinitely many times on a small interval (see the comments preceding proposition A.1.10 for a full discussion). This is another (potentially significant) complication which Jones fails to mention. It can be got round however by hitting γ with a small perturbation, nudging its diagram so that

D_γ ends up only having simple zeroes, which avoids the possibility of these infinitely wiggly bits. The proof of this result (proposition A.1.10) is the first that requires a real idea, in this that of introducing a small fairly arbitrary perturbation to γ, and proving that some such small perturbation will have the required result.

We introduce the terminology that γ **projects nicely** if it has regular projection avoiding 0, with D_γ only having simple zeroes – thus ruling out the possibility of infinitely many wiggles just discussed. Then with the results mentioned here proved, the key result that needs to be established is the following.

Bending Forward proposition. *Let γ be a smooth knot which projects nicely. Let U be an open interval in $\mathbb{R}$ such that γ has at most one backwards bend on U. Then there is a smooth knot β which projects nicely and is a bending forwards of γ such that if $t \in U$ then $D_\beta(t) > 0$.*

If this proposition can be shown, Alexander's lemma follows easily, as can be seen in lemma A.1.24 and the immediately preceding results. The question now is what we have to do to rigorously establish this proposition.

6 The "over the shoulder" manoeuvre

To argue that if γ has at most one backwards bend on U then we can bend it forwards in this way, Jones's account is simple. We split U up into short sections on which it has at most one crossing, and throw each of these over our shoulder (if our stretch lies above the part it crosses), or throw it the other way (if our stretch lies below the part it crosses). As discussed in section III.4, Jones misses out a key part of the argument here, which is that if we have divided U up into short stretches $V_1, \ldots V_n$, then we need to make sure that when throwing V_i over our shoulder we don't add any crossings to any unfixed V_j. This is illustrated in fig. III.3.

This is the part of the proof that Jones thinks would be very difficult to argue in more detail: he says that

> constructing the functions required for the "throwing over the shoulder" trick would be a nightmare. (Jones 1998, p. 212)

It is not clear what method he envisages that would be so nightmarish. It is true that it would probably be extremely difficult to concoct a single formula for the path in fig. III.3 directly built from basic smooth functions. The same goes for the smooth isotopy. It may not even be at all clear whether a suitable formula for a smooth isotopy exists; maybe there is a worry one would have to resort to nudging little bits of the curve in the right direction, one small bit moving one small step at a time. This could well be a tremendous effort.

Nonetheless, a proposition being hard to prove is not generally grounds for omitting to prove it, when doing rigorous mathematics. If a proposition seems obvious but a proof seems very hard, that suggests that either the proposition is not as obvious as it appears (as with the Jordan curve theorem), or that our proposed tactics for proof are lacking and we should look for better ones.

This goes even for propositions which may be very intuitive: there are plenty of examples of such propositions that are normally taken to require careful proof. The separating hyperplane theorem states (in one version) that if A and B are disjoint compact convex subsets of $\mathbb{R}^n$ then there is a hyperplane which separates them, lying strictly between them. Here a convex body is one containing the line segment between any two points in it, and a hyperplane in $\mathbb{R}^n$ is an $(n-1)$-dimensional affine subspace. This theorem is illustrated in fig. III.4. In dimensions 2 and 3 it is about as intuitive a statement as one can imagine. Convex bodies "look" very straightforward, and there is no way to draw or visualize two such compact convex bodies without the statement seeming completely obvious from an intuitive perspective. Nonetheless proving this takes some thought (give it a try).

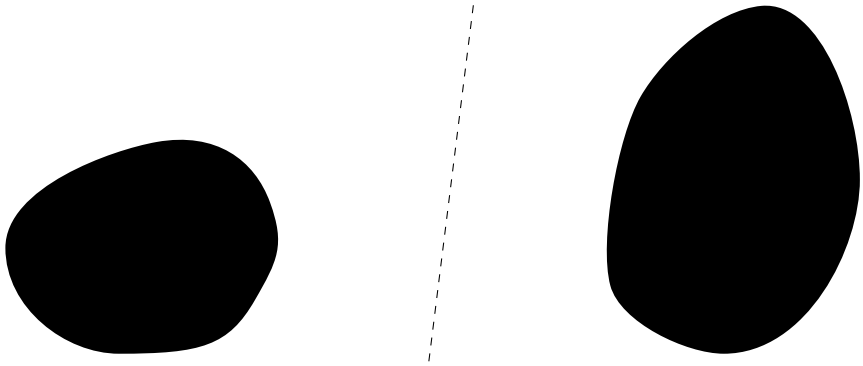


Figure III.4: The separating hyperplane theorem

I am not aware of anyone seriously suggesting a rigorous argument could assume a proposition like this without proof, no matter how intuitive it is. Perhaps in a more applied area of mathematics one might do so, if the intuition was strong enough that one found it convincing; but as discussed in section I.2, reasoning can be convincing and reliable without being rigorous. One of the normal purposes of the label "rigour" is to make clear that if propositions such these are being relied on, then they will have been proved.

If someone wishes to claim that a rigorous argument could assume the Bending Forward proposition of section III.5 without proof, then they have two options: arguing that rigour does not actually need propositions like the above to be proved, or arguing that the case of the Bending Forward proposition is somehow different. The former amounts to saying that actually there was no need to prove the Jordan curve theorem, or the separating hyperplane theorem, which is not an option I imagine many will find attractive. It is also not obvious how an argument for the latter option would proceed. If anything, the separating hyperplane theorem (in dimensions 2 and 3) is *more* immediate from an intuitive point of view than the Bending Forward proposition. Perhaps one could say that the separating hyperplane theorem is likely to have wider application, so we need to justify it more surely; but that is an argument for enforcing rigour in one case and not the other, not an argument that what rigour requires is different in the two

cases.

The requirement that we can provide a proof of the Bending Forward proposition is an example of the general requirement from chapter I that we be able to prove inferences in greater detail. In this case as it stands, the Bending Forward proposition is too coarse an inference (though plausible) to be acceptable in a proof by the usual standard of rigour: though it can be proved, even filling in its details at all requires a decent amount of ingenuity and effort. By providing a more detailed justification of it, we will reach the point where (though not completely formal) making any point in the argument yet more detailed becomes a routine matter.

So the question is how we can justify this over the shoulder manoeuvre. Here we can see more clearly what is going on by splitting our task up into two separate claims. One is the existence of a suitable path in the plane like the dashed path traced out in fig. III.3. The second is the existence of a smooth isotopy modifying our original knot to follow this new path (with its planar projection). We will deal with this latter aspect first – the part that Jones thinks will be a "nightmare".

Part of carrying out the smooth isotopy can be justified very easily. The point of dealing with a short stretch of knot on which the diagram has at most one crossing is that then we can pull this stretch far out of the page, or push it far into the page (depending on whether our stretch lies above or below the stretch it crosses), taking it away from the main body of the knot so that we can then manoeuvre it freely. We can obtain the following precise result to this effect, whose proof is completely routine.

Proposition III.7. *Let γ be a smooth knot such that $\gamma'_{\mathbb{C}}(t) \neq 0$ for all t, and $\gamma_{\mathbb{C}}$ has at most one crossing point in (a, b), with $a < b < a + T$. Let $K > 0$ such that $|\gamma_3(t)| < K$ for all t, and let $a < c < e < f < d < b$. Then γ is smoothly isotopic to a smooth knot β with:*

- $\beta_{\mathbb{C}} = \gamma_{\mathbb{C}}$

- $\beta(t) = \gamma(t)$ *for* $t \notin (a,b) + T\mathbb{Z}$
- $|\beta_3(t)| \leqslant K$ *for* $t \notin (c,d) + T\mathbb{Z}$
- $|\beta_3(t)| \geqslant K$ *for* $t \in [c,d]$
- $|\beta_3(t)| = 2K$ *for* $t \in [e,f]$
- $(|\beta_3|)'(t) > 0$ *for* $t \in (c,e)$
- $(|\beta_3|)'(t) < 0$ *for* $t \in (f,d)$

Thus our modified knot β has the same projection as γ, and has the same "vertical" values as γ too outside of $(a,b)+T\mathbb{Z}$. $\beta(t)$ flies steeply up from the plane for t just above a, reaching a (potentially very great) distance K from the plane for t between c and d, and then $\beta(t)$ heads steeply back down to the plane as t approached b. We can take the intervals (a,c) and (f,b) on which β is heading up from the plane and then down again to be as small as we like, and can take the distance K to be as large as we like.

Having pushed this section of the knot far above (or below) the rest of it, we are free to slide and bend it around as we like without hitting the main body of the knot. When we think about it, we have great freedom to slide curves around in a given plane Consider fig. III.5 for instance. It seems obvious that we can push the curve over to follow the new dashed path, without moving A or B: just push the solid section between A and B down through the gap, then stretch out the left hand side, bringing it past the left of A to trace out the dashed path. While doing so the tangent to the curve at A will rotate in a circle, but that will not be a problem (since we can take the curve to be bending down very steeply towards the plane near A).

This manoeuvre does not depend on any special features of the curves pictured. Even if the target curve is very wiggly, that makes no difference. Consider the situation in fig. III.6: perhaps it is harder to see in this case, but again we can isotopy to follow the

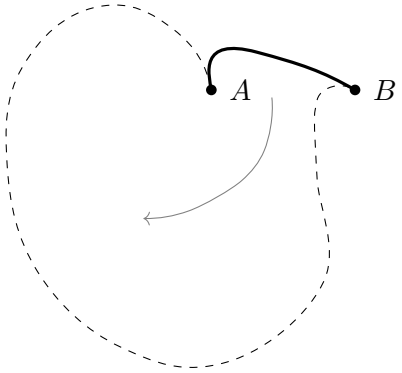


Figure III.5: A planar isotopy

dashed path with no problem, pushing it down through the gap between A and B and stretching and sliding parts of the curve into the appropriate places.

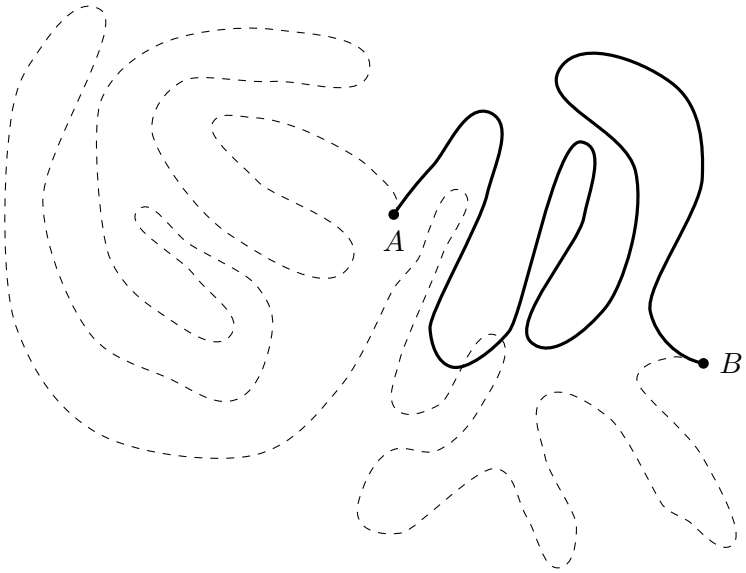


Figure III.6: More complicated planar isotopy

Considering examples like this leads us to suspect that actually any two smooth planar curves with the same endpoints will be smoothly isotopic (by an isotopy which fixes their endpoints). This may seem like a foolish point to raise – why would we consider general planar curves, when we could limit our attention to simpler examples like that in fig. III.5? But even in that simple case it was not clear how a proof should

go, and as Jones says it might be a "nightmare" to try to construct a smooth isotopy by hand.

It is common in mathematics that theorems are easier to prove when proved in the right generality, which can mean proving them in more generality. If it is the case that any two smooth planar curves with the same endpoints are smoothly isotopic (keeping endpoints fixed), this is in a sense a more natural fact than a special case like fig. III.5, and maybe we would expect there to be a good "reason" for this. Perhaps one might worry that proving the general case will just be a harder version of the simpler case, maybe constructing an isotopy to push in each wiggly part one at a time, giving an inductive argument on the number of wiggles. But the point of considering the general case is to lead us away from this kind of thinking.

As it happens, there is a very neat proof of the general case. It may well be that this is a standard result, but I have not been able to find a reference for it. The key idea is that a smooth wiggly path gets less and less wiggly as we zoom in on it – that is, shorter and shorter sections of the path get less and less wiggly. Thus if we zoom in to shorter segments of the path, whilst simultaneously rescaling to keep the endpoints fixed, we will follow a smooth progression that transforms our path into a line. For instance zooming in on the midpoint of the dashed path in fig. III.6 and rescaling gives straighter and straighter sections, as seen in fig. III.7. Since being smoothly isotopic (relative endpoints) is an equivalence relation, if every such path is smoothly isotopic to a straight line in this way then we are done.

Using this idea we can indeed prove that any two smooth planar curves with the same endpoints are isotopic (keeping the endpoints fixed). Actually it is convenient to prove a slightly stronger statement, in which the endpoints can vary within some complex neighbourhood at each end, introducing the notion of a **smooth isotopy relative** $(a \hookrightarrow X)$ for a smooth isotopy in which the value of the curve at endpoint a stays within the convex set X throughout the transformation (see the discussion preceding

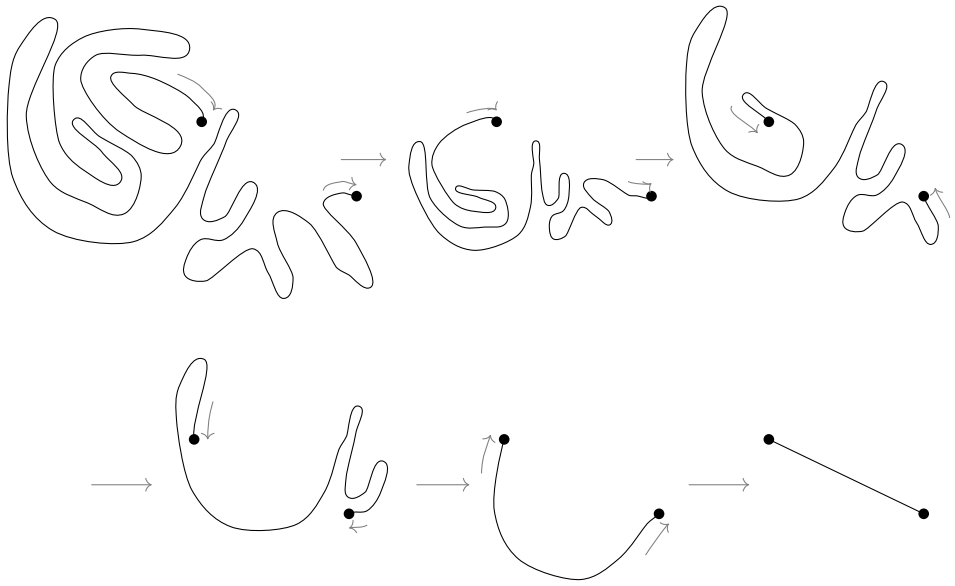


Figure III.7: Zooming in on a wiggly path

proposition A.1.12 for a more formal statement). The result is the following. Here a smooth immersion $\gamma : \mathbb{R} \to \mathbb{R}^2$ is a curve with $\gamma'(t) \neq 0$ everywhere.

Proposition III.8. *Let I be a proper interval and let $\alpha, \beta : I \to \mathbb{R}^2$ be injective smooth immersions. Let $a \neq b \in I$ and let X, Y be disjoint convex subsets of $\mathbb{R}^2$ with $\alpha(a), \beta(a) \in X$ and $\alpha(b), \beta(b) \in Y$. Then there is a smooth isotopy relative $(a \hookrightarrow X)$ and relative $(b \hookrightarrow Y)$ from α to β.*

The proof is elegant, with most of the work just checking that the isotopy sketched in fig. III.7 is smooth (which is not particularly difficult). This kind of result could well be useful in other circumstances as well, and the theorem has obvious higher dimensional analogues.

With this and proposition III.7, it is not difficult to justify a very general form of the over the shoulder manoeuvre (there is a routine intermediate result proved along the way in section A.1, which we skip here).

Proposition III.10. *Let γ be a smooth knot which has regular projection, with $a < b <$*

$a+T$ such that $\gamma_{\mathbb{C}}(a) \neq \gamma_{\mathbb{C}}(b)$ and $\gamma_{\mathbb{C}}$ has at most one crossing point in (a,b). Suppose $\alpha \in C^{\infty}_{T}(\mathbb{R},\mathbb{C})$ is a smooth immersion such that $\alpha|_{[a,b]}$ is injective, and with $\alpha(t) = \gamma_{\mathbb{C}}(t)$ for $t \notin (a,b)+T\mathbb{Z}$. Then γ is smoothly isotopic to a smooth knot β with $\beta(t) = \gamma(t)$ for $\notin (a,b)+T\mathbb{Z}$, and $\beta_{\mathbb{C}} = \alpha$.

This is the result that Jones claims would be a "nightmare" to prove. In this he is simply wrong. The proof is largely routine, with the only little bit of magic needed being the idea behind proposition III.8. By skipping over the proof Jones misses out on a chance to discover that neat idea for himself.

The last step needed to complete the argument is the existence of a suitable target path. We need to argue that given a smooth knot, a planar path like that in fig. III.5 exists; if we can do this then we can use proposition III.10 to deduce the existence of a suitable smooth knot, one which is a bending forwards of our original knot.

It will not be possible to cover this last part in detail. It does not require any great insight, though is a certain amount of effort. The tactic used in section A.1 is to first prove the existence of a suitable angle function, then join this with a suitable radius function, and then perturb the result to make it regular. The resulting proposition is complicated (proposition A.1.17), as the resulting knot needs to satisfy a number of different conditions to allow the Bending Forward proposition to be proved. This complexity is no reason to sweep the details under the rug though – it suggests that someone who claims the existence of such a knot is obvious has not fully grasped what is required.

With this in hand, finishing the proof is largely straightforward. Finally we obtain our desired conclusion.

Lemma III.24 (Alexander's lemma)**.** *Let γ be a smooth knot. Then there is a smooth knot β which is smoothly isotopic to γ such that β has regular projection avoiding 0, and we have $D_{\beta}(t) > 0$ for all t.*

7 Assessment of Jones's argument

As we have seen, there are numerous major problems with Jones's argument from a rigorous point of view. The basic structure of the argument is very unclear, with him skipping over a key aspect, one which substantially complicates it (section III.4). The terms he uses – like "short stretch" – are also very unclear, and properly defining them is a delicate task (section III.5). Then, with this preliminary work done, the main result still needs to be proved. Section III.6 looks at similar mathematical results which are taken to require rigorous proof, and argues that this is no different: there is no reason it should be regarded as rigorous to just assume this result, or gesture at a picture as an argument. Section III.6 sketches the argument, which employs a neat idea at one point to prove that any two smooth planar curves are smoothly isotopic relative their endpoints.

When this work is done, the resulting argument is rigorous by the normal standard, as described in chapter I: all its concepts are defined precisely, and all its inferences are written out at a level of pretty great detail, and each one could be proved in even greater detail if one desired. Thus Jones's argument is no threat to the view of rigour put forward in chapter I. The comments of Jones, De Toffoli and Giardino, and Larvor – about the argument involving essentially visual reasoning and high level intuition, that would be very difficult to justify in more detail or formalize – are only true of Jones's non rigorous version of the argument. Making it rigorous removes these features.

The result is much longer and more complex than Jones's simple sketch, as seen in section A.1. However, we can see some definite benefits of making Jones's argument rigorous. Not only is the argument much clearer and more explicit, but the process generates potentially useful mathematical ideas. As mentioned in section III.5 and discussed fully in section A.1 (proposition A.1.10 and the preceding), the proof that the set of knots which project nicely and for which D_γ has only simple zeroes uses a perturbation

argument, proving that a certain modification of a knot exists by considering a family of small perturbations of the original knot, and proving that one has the required property. This is an idea which could obviously be useful more widely in the subject and in other areas of mathematics (and I'm sure is a standard technique in some areas). Then the proof of proposition III.8 involves another new idea – the zooming and scaling idea – which has aesthetic value in itself, as well as straightforward generalizations. Also the result proved could well be useful in other contexts. Much of the argument of section A.1 does consist of filling in routine details, and may not have much relevance beyond this particular result, but such is mathematics.

8 Rigorous use of pictures II

Jones's argument can be used to illustrate a wider point about the use of pictures in mathematics, as discussed in sections III.1 and III.5. This is that for a mathematical argument involving pictures to be rigorous, we have to be able to state which features of the pictures actually play a part in the argument, and which are merely artefacts of how it happens to be drawn – we have to know which features of the pictures are argument relevant.

For instance, if a section of the diagram is roughly straight, are we intending to argue as though it is actually roughly straight, or is that just how we drew it? Actually we need to know much than just whether the section is being represented as "roughly straight". If we are going to rigorously reason with the diagram, we need to know which clearly stateable property the diagram is indicating – for instance, a certain precise bound on the curvature of the curve in that area, or that the curve has no parametric inflection points. This is what I mean by knowing which features of the picture is argument relevant; knowing which precisely stateable properties it is representing. Of course a picture may just be indicating such a feature, rather than sharing it exactly.

In section III.5 we saw how unclear Jones's notion of "short stretch" was: there are any number of properties that we might think one of his "short stretches" should have, and we have no clue which of these are actually required by the argument. A huge variety of such properties are compatible with the diagrams Jones gives. The same goes for the over the shoulder manoeuvre – we do not know what properties the result is required to have, as a wide range of properties are all compatible with Jones's diagrams. Thus because we do not know which features of Jones's diagrams are actually playing a role in the argument, and which are accidental, we cannot regard the argument as rigorous: we do not know what key terms in the argument mean[6], and do not know what inferences we are expected to be making.

This lesson applies much more widely. Suppose we have a diagram A and a diagram B, and we are asked to assent that from A we can reach B. If there is any generality involved, and we do not know what general features of the two diagrams are included, then we simply do not have a rigorous understanding of what we are being asked.

This is an important point, since it defuses the threat that pictorial reasoning has been thought to present to formalizability, and to the kind of account of rigour put forward in chapter I. It also constitutes an amendment to Larvor's account of what pictorial arguments have to satisfy to be rigorous (Larvor 2019).

First, the objections to formalizability. As mentioned in the introduction, many of those who object to formalizability as a norm in mathematics focus their objections on pictorial arguments in particular (Leitgeb 2009; Goethe and Friend 2010; De Toffoli and Giardino 2015; 2016; Larvor 2019). They claim that pictorial arguments may involve reasoning that resists verbal description. But if a pictorial argument is rigorous by the above standard, we see that this cannot be the case.

Suppose we have an argument formed from a sequence of pictures, and that the

[6]In the mathematical sense: of course we may know what to provide if a "short stretch" of string is requested, but this is no more mathematical use than an informal every day notion of continuity would be to analysis.

argument is rigorous – so we can state which features of the pictures are actually relevant to the argument, and clearly describe these features. Then (us being finite beings) each picture will only have finitely many such relevant features, so we can replace each picture by a description of the relevant features. As discussed initially, we need to be able to make this description mathematically clear for our use of the picture to be rigorous. Thus we should be able to straightforwardly convert our pictorial argument into a sequence of clear textual inferences. A reader is free to visualize the meaning of these textual inferences, as with any other textual inferences, and they may even visualize something like the omitted pictures.

In a sense this process is illustrated by the process of rigorizing Jones's argument seen in sections III.4 to III.6, but there is a lot going on there beyond just converting Jones's pictures into prose (due to the many different respects in which his argument lacks rigour).

One objection might be that if we replace a sequence of pictures by a sequence of blocks of text in this way, the result might be very long and complex. In many cases we would not expect much increase in proof length though – if a sequence of pictures comes with a statement of which features are argument relevant, then ditching the pictures and just using this statement instead might even reduce the page count of the argument. The textual proof will only be much more cumbersome if we have some compressed way of conveying information in the pictures, a convention for which features of the pictures are actually argument relevant. Such a convention is illustrated by the case of Euclidean diagrams, as spelled out by Mumma (2010). With the convention that only coexact features of Euclidean diagrams are actually argument relevant, it is possible to give efficient diagrammatic proofs, that may (perhaps) be much longer if written out in prose.

This objection ultimately bleeds into the general objection often raised to formalizability, that filling in the details of an argument may fundamentally alter it, overwhelming

the reader with trivialities and obscuring the bigger picture. I cannot reply to that objection here, but my point is that it is a completely general objection, and nothing specific to the case of diagrammatic reasoning. Pictures are just one possible example of an efficient high level way of conveying information, one amongst many others. High level notation does not have to be pictorial. If the use of pictures is rigorous in the sense defended here – if we know which features are argument relevant – then the debate about formalizing pictorial arguments is much like the debate about formalizing any other kind of argument.

As well as defusing these picture based objections to formalizability, this condition – that we can state which features of diagrams are argument relevant – constitutes an amendment to Larvor's account of pictorial rigour (Larvor 2019). Larvor states three necessary conditions that diagrams must satisfy to be rigorous:

(a) it is easy to draw a diagram that shares or otherwise indicates the structure of the mathematical object

(b) the information thus displayed is not metrical

(c) it is possible to put the inferences into systematic mathematical relation with other mathematical inferential practices.

He does not claim that these are jointly sufficient, instead aiming to give necessary conditions that can be supplemented in future work (ibid., p. 6). This I aim to do.

Larvor motivates these by discussing various case studies, including Euclidean geometry and Jones's argument. As we have seen, Jones's argument is not rigorous, so should not be used as the basis for a philosophical study of rigorous pictorial arguments, but Larvor's conclusions do not rely on it.

I do not think that these conditions are incorrect, but I do think there is more to say. The most important point is the point made above: that we know which features of a diagram are argument relevant (which may well depend on what argument we are

using it in). This I think makes condition (c) more or less redundant – if we know which features of a diagram are argument relevant, then we already know how the diagram should interact with other inferential practices. We can only rely on those features of the diagram that we were taking to be argument relevant when drawing new inferences from it. We cannot rely on any features of the diagram that were merely artefacts of how it happened to be drawn. Similarly if we are to infer a diagram from some other inferential context, then we have to be sure that all argument relevant features of the diagram can actually be inferred from that other context. If these conditions are satisfied then there seems to be no barrier to rigorously reasoning between the diagram and other contexts.

This also clarifies (a). When drawing a diagram that shares or otherwise indicates the structure of the mathematical object, if the diagram is going to be rigorous then we have to be clear on which of its features are intended to be part of the representation. I have no quibble with (b). Thus we obtain the following amended version:

(a) it is easy to draw a diagram that shares or otherwise indicates understood features of the mathematical object

(b) the information thus displayed is not metrical

(c) we are able to clearly state which features of the diagram we are taking to be part of its role in an argument, and which are merely accidental

Here we need to have a mathematically precise understanding of the "features" of a diagram, as discussed above.

With these modifications made, we have a simpler, clearer and stronger set of necessary conditions pictorial arguments in mathematics to be rigorous. We also defuse worries about whether pictorial arguments might be unformalizable.

Chapter IV

Ancestrals, Primitive Recursion and Isaacson's Thesis

For the next two brief chapters we shift focus, to address some conceptual issues that will be appealed to in the discussion in chapter VI of what we want from a foundation for mathematics. The primary question investigated in this chapter concerns primitive recursion: what logical resources are required to capture our grasp of primitive recursion, and the ability to define primitive recursive functions out of a simply infinite sequence?[1] The answer given here will be in the form of an operator which I call the double ancestral, a version of the ancestral operator with two arguments instead of one (or four instead of two, depending on how one conceives of the ancestral). This is not an entirely novel insight: R. M. Martin (1943; 1949) defined a version of the ancestral slightly more general than that given here and showed how it could be used to define primitive recursive functions. He did not advocate this as the real conceptual basis of primitive recursion however, or discuss the philosophical implications of this for arithmetic, mainly being concerned with developing a relatively strong nominalistic framework.

[1]By simply infinite sequence I mean the usual notion of ω-length infinite sequence equipped with successor function, such as the natural numbers structure.

Since the double ancestral operator is arguably a purely logical operator, and is ontologically innocent, we thus obtain that the ability to define primitive recursive functions out of a simply infinite sequence is available in purely logical terms. This will be used in chapter VI, to show that one can interpret arithmetic statements in terms of any real simply infinite sequence that we encounter – part of the more general discussion of real world instantiations of mathematical structures, which is important to the question of soundness phrased in chapter VI.

The conceptual issues considered in this chapter and the next give an opportunity to do some philosophy of a rather different flavour, and as an aside, this chapter will discuss implications of this issue for Isaacson's thesis (this is an incidental application of independent interest, and not appealed to in the rest of the thesis). This is a thesis concerning the status of PA (first order Peano arithmetic). Though PA is necessarily incomplete, Isaacson (1987; 1992) famously argues that there is a sense in which it is complete: it captures the purely arithmetical content of our concept of natural number. Isaacson's thesis states that to prove an arithmetical sentence which is unprovable in PA, one will have to employ further ideas, such as higher order concepts or reflections on the consistency or truth of the axioms of PA. These further ideas go beyond the purely arithmetic.

Isaacson argues mainly by looking at examples of true sentences unprovable in PA, and seeing what is needed to prove them. Smith (2008) gives a different argument for the same basic thesis, arguing that understanding the predicate "natural number" amounts to understanding the ancestral operator — and thus that the truth of Isaacon's thesis rests on whether when you supplement PA with the ancestral operator in the appropriate way, the result is conservative over PA. As Smith demonstrates, it is not difficult to show that it is, giving positive support to Isaacson's thesis.

This chapter considers the case of primitive recursive functions, and uses this to buttress Smith's argument. Indeed the same questions Smith asks of the predicate

"natural number" should be asked of the functions of addition and multiplication. In Peano arithmetic these functions are assigned symbols in the language, governed by the axioms

$$x + 0 = x$$

$$x + Sy = S(x + y)$$

$$x \times 0 = 0$$

$$x \times Sy = (x \times y) + x$$

where S is the successor operation on numbers. Just as we can ask how we form the predicate "natural number" and know its axiomatization in PA is appropriate, we can ask this of $+$ and $\times$. We are not simply positing these functions, assuming that there are valid operations on numbers with these properties. We are not imposing these operations by fiat – perhaps starting with multiple candidate infinite sequences, and then narrowing our attention to those which happen to allow these operations. We feel we can see that numbers are the kind of things which can be added and multiplied in the way these axioms describe. Grasping this is part of grasping the axioms of PA. But how do we grasp this? How do we come to see that we can introduce functions like this, which are total and single and satisfy the relevant equations? As we will see, this question has a satisfying answer in the form of a double version of the ancestral operator.

Since the double ancestral provides a plausible and satisfying account of our grasp of these primitive recursive functions, we obtain a new test of Isaacson's thesis: when arithmetic is phrased in terms of the double ancestral, is the resulting theory conservative over PA? This is a stronger test than Smith's since this theory straightforwardly interprets the ancestral arithmetic used in his test.

1 The thesis

We start by clarifying what Isaacson's thesis states. In Isaacson's original argument, he takes the proper conception of the natural numbers to be given by the second order axiomatization. He then feels it is natural to see what remains of this when one restricts oneself to just quantifiers over the natural numbers (rather than also sets of natural numbers), moving from the second order induction axiom to the first order induction axiom scheme. Since what one obtains from this is too weak to define addition and multiplication, axioms for these are added (and all primitive recursive functions then become definable). The theory that results from this is PA (Isaacson 1987, pp.148–154). For Isaacson, Isaacson's thesis is that any arithmetic truth is provable in PA, where an arithmetic truth is one which is part of the "purely arithmetic content of our full understanding of the concept of natural number". The latter seems to be shorthand for the process described above by which PA arises.

Thus stated, the thesis appears to be true, but somewhat ad hoc. If our concept of natural number is a second order concept, then why should "arithmetic content" refer to restricting oneself to first order quantifiers over numbers? Further, the above characterization of "arithmetic truth" seems to be in conflict with his later remarks about it. He talks of arithmetic truths being those that can be directly perceived to be true, or derived from such statements (ibid., pp.159, 160, 162, 163, 165, 165, etc); but if grasping the concept of natural number really means grasping the second order axiomatization, then the first order instances of the induction scheme are not directly perceived to be true, they are deduced from the second order axiom. Similarly if addition and multiplication are rightly defined in the way Isaacson discusses, using Dedekind's method which quantifies over functions as objects, then one does not directly perceive the Peano axioms for them to be true – these axioms have a fairly complicated second order justification. Given that these statements can be seen to be true on the basis of

the second order axiomatization, there seems to be no principled reason why could not give a second order justification for further first order arithmetic truths. That would tell against Isaacson's thesis.

Nonetheless, one can extract from Isaacson's writings the following central idea (here we let L_A be the language of PA – a first order language):

> Proving a statement of L_A that is unprovable in PA will require employing concepts beyond those required to grasp the basic concepts of arithmetic: natural number, successor, induction, addition and multiplication.

This is what I mean by Isaacson's thesis in this chapter.

How strong a thesis this is will depend on how strong a notion of "grasp" one works with – to fully understand a certain concept, how much further one's understanding should extend. In general these questions can be difficult. Sometimes there is a consensus, as in the generally held view that one can properly grasp first order logic without being able to understand second order logic. Things are not always as clear as in this case though – can one fully grasp fifth order logic without being able to grasp sixth order logic?

What the thesis amounts to may also depend on what one takes the right interpretation of arithmetic to be. For instance suppose one defended a conception of arithmetic as being about strings formed from some particular symbol |. This is essentially the conception of Hilbert (1990), and is developed in more detail by Parsons (2007), though they are concerned particularly with intuitive aspects of the theory and avoid arbitrary quantification over the domain. It may be that one could describe a first order theory of the strings, interpret PA in this theory, and argue that our grasp of the concepts of arithmetic amounted to a grasp of this theory of strings (this is not what the authors mentioned argue). Then if one could show that this theory of strings was conservative over PA, one would have evidence for one version of Isaacson's thesis. However this would be a very limited version of Isaacson's thesis, entirely dependent on the claim

that a proper grasp of arithmetic amounts to a grasp of this theory of strings and should be expected to extend no further. It would be more an argument for a particular interpretation of arithmetic than for Isaacson's thesis in general.

The best defence of Isaacson's thesis would be one which examines the axioms of PA themselves, rather than relying on any particular interpretation of them. One will also obtain a stronger version of Isaacson's thesis if one is liberal in the notion of "grasp" one uses – liberal in questions of what further concepts proper grasp of a particular concept entails.

2 The argument

Smith (2008) gives a better argument for the thesis than the hypothetical string based one just sketched. He focuses on what is required to grasp the concept of natural number. The basic thought is that

> understanding quantification over the [natural] numbers involves understanding that the numbers are zero, the next number, the one after that, *and so on*, *without limit* – and understanding too that these are the *only* numbers. Which is in effect to grasp the thought that every number stands in the ancestral of the successor relation to zero. (ibid., pp.3–4, emphasis his)

Smith thus argues that grasping the concept "natural number" amounts to grasping the ancestral operator. This allows him to set a test for Isaacson's thesis. He supplements PA with the ancestral operator, to give what he calls "ancestral arithmetic". Then if grasping the concept "natural number" amounts to grasping the ancestral operator, Isaacson's thesis requires that anything provable in this ancestral arithmetic is already provable in PA, i.e. that ancestral arithmetic is conservative over PA. This Smith shows straightforwardly, giving positive support to Isaacson's thesis.

However Smith's account misses out a crucial part of arithmetic, as discussed initially:

the functions of addition and multiplication. We want to know how we grasp that numbers are the kinds of things that can be added and multiplied. One approach is to define addition and multiplication in full second order logic, which one can do given the successor operation and a second order induction axiom. It would be a surprise if quantifying over relations was necessary to grasp these primitive recursive functions however. If true, that would presumably disprove Isaacson's thesis as understood here.

It would also suggest that primitive recursion requires the existence of abstract objects, so is not a purely logical operation, and is unavailable to a nominalist; although a nominalist will not be discussing addition and multiplication of numbers, there may be other contexts where they wish to use primitive recursion, for instance involving concretely instantiated infinite sequences. If understanding primitive recursion required quantifying over relations, that looks impossible however.

The main claim of this chapter is that the double ancestral gives us a satisfying analysis of how one can grasp these kinds of primitive recursive functions, in the same way the ancestral does for the concept of "natural number". This gives rise to a new, stronger test of Isaacson's thesis, in terms of what I call double ancestral arithmetic.

One issue I will not address is the Neo-Fregean analysis of arithmetic. They might argue that arithmetic is properly understood in terms of cardinality, using second order logic augmented with Hume's principle. If that were true it would present a major challenge to Isaacson's thesis as understood here. I do not find the Neo-Fregean arguments convincing, but they are not the subject of this chapter, and will have to be set to one side.

3 The ancestral and the double ancestral

The prototypical instance of the ancestral operator is the relation "ancestor". Similarly a prototypical example of the double ancestral operator is the relation "ancestor of the

same generation". Figure IV.1 illustrates this diagrammatically: illustrating the relation $\mathrm{ASG}(x, y)$ between ancestors of Jeff and ancestors of Sarah, of x being an ancestor of Jeff of the same generation as y is an ancestor of Sarah.

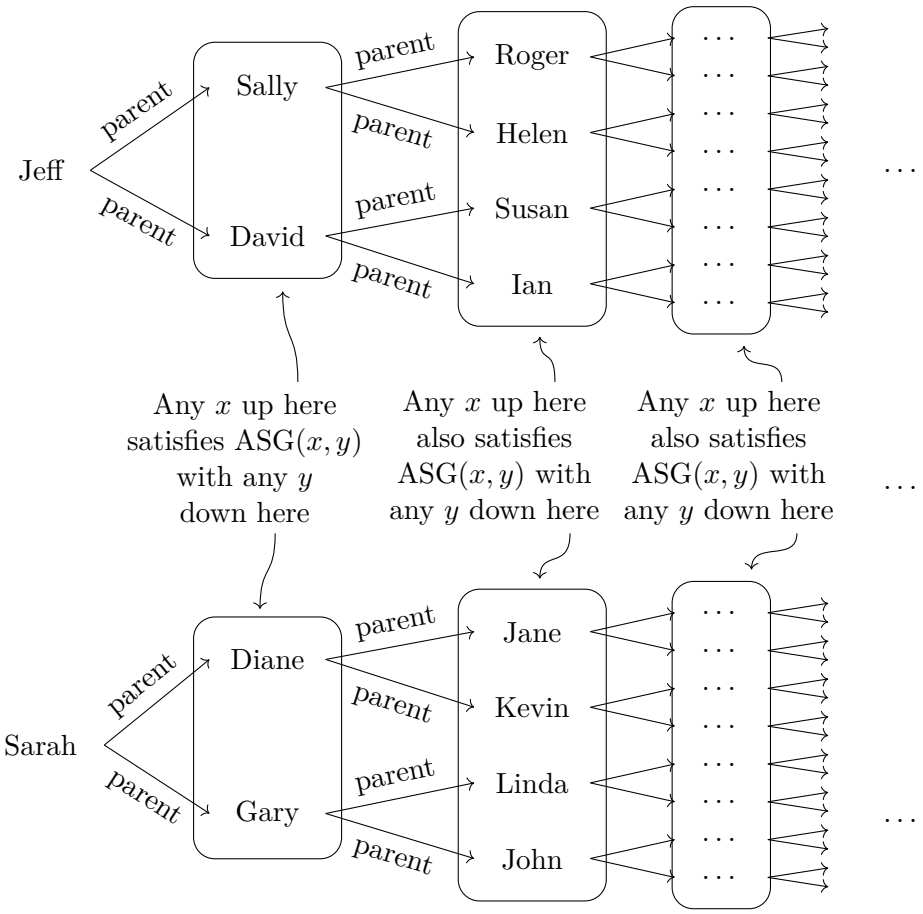


Figure IV.1: The "ancestor of the same generation" relation

We can illustrate the general case of the double ancestral operator in the same way, seen in fig. IV.2. To better suit its application to primitive recursive functions, we will use the reflexive form – this corresponds to a modification of the "ancestor of the same generation" relation to include Jeff as an ancestor of Jeff of the same generation as Sarah is of Sarah. We let $\phi(x, y)$ and $\psi(w, z)$ be two place relations, and write $(\phi, \psi)^*(c, d, x, y)$ to indicate that the double ancestral of ϕ and ψ holds of c, d, x and y. In fig. IV.2 we

use $a \xrightarrow{\phi} b$ to indicate that $\phi(a, b)$ holds, $a \xrightarrow{\psi} b$ similarly.

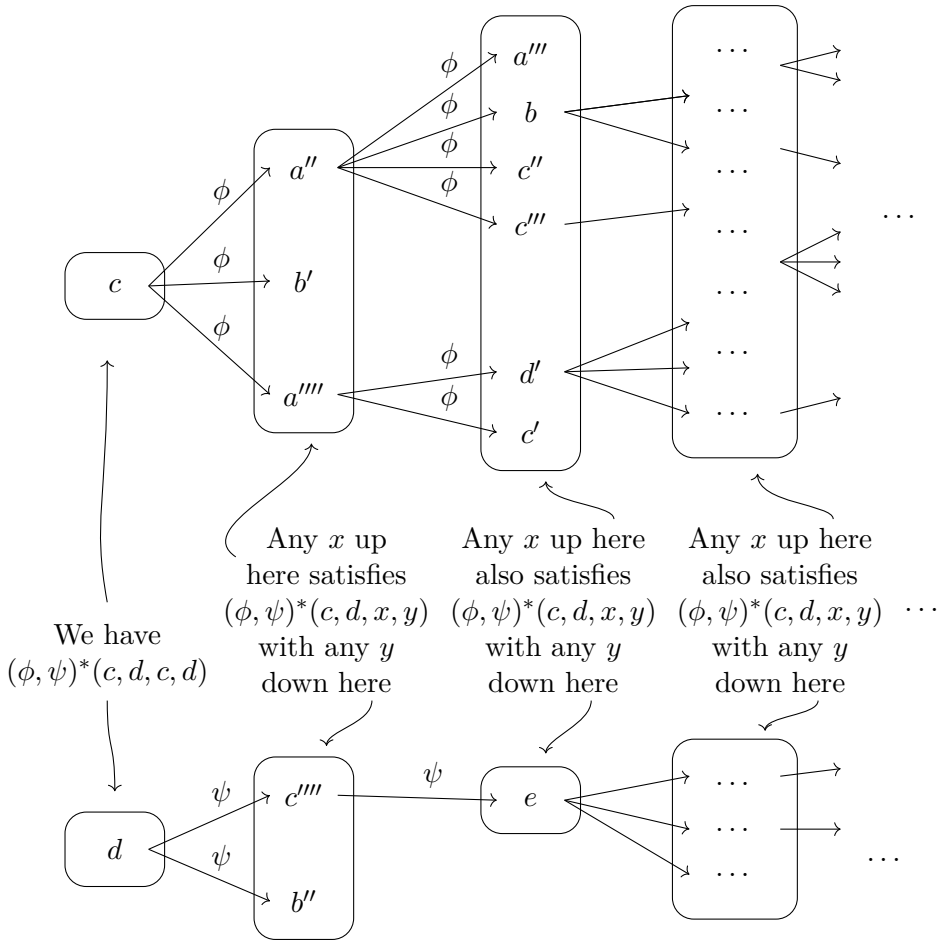


Figure IV.2: The double ancestral of ϕ and ψ

We can also informally explain $(\phi, \psi)^*$ in prose. We have that the relation $(\phi, \psi)^*(c, d, x, y)$ holds iff

- $x = c$ and $y = d$
- Or $\phi(c, x)$ and $\psi(d, y)$
- Or there are u and v such that $\phi(c, u)$ and $\phi(u, x)$, and $\psi(d, v)$ and $\psi(v, y)$
- Or there are u, u' and v, v' such that $\phi(c, u)$ and $\phi(u, u')$ and $\phi(u', x)$, and $\psi(d, v)$ and $\psi(v, v')$ and $\psi(v', y)$

- Or there are u, u', u'' and v, v', v'' such that $\phi(c, u)$ and $\phi(u, u')$ and $\phi(u', u'')$ and $\phi(u'', x)$, and $\psi(d, v)$ and $\psi(v, v')$ and $\psi(v', v'')$ and $\psi(v'', y)$

$$\vdots$$

and so on (and these are the only objects related by $(\phi, \psi)^*$). This is strictly analogous to how one would informally explain the ancestral, except that it is a simultaneous description involving two relations ϕ and ψ rather than just one – if you left out all mention of d, y and ψ from the above you would have a description of what is required for the ancestral $\phi^*(c, x)$ to hold.

One can give a precise definition of the double ancestral using finite sequences. We have that $(\phi, \psi)^*(c, d, x, y)$ holds iff for some $n \geqslant 0$ we have sequences $(a_0, \ldots a_n)$ and $(b_0 \ldots b_n)$ such that $c = a_0$, $d = b_0$, $x = a_n$, $y = b_n$, and for all $i = 0 \ldots (n-1)$ we have $\phi(a_i, a_{i+1})$ and $\psi(b_i, b_{i+1})$. One can also give a definition in second order logic, where the relation $\{(x, y) \mid (\phi, \psi)^*(c, d, x, y)\}$ is the intersection of all relations R such that $R(c, d)$ and such that if $R(x, y)$ and $\phi(x, w)$ and $\psi(y, z)$ then $R(w, z)$.

However there seems no reason to think that understanding predicates formed from the double ancestral operator *requires* one of these definitions – any more than for the ancestral operator. Smith (2008) and Avron (2003) argue that an explicit definition of the ancestral is not necessary, and that the ancestral operator can be thought of as a conceptual primitive occupying a valid middle ground between first and second order logic.[2] Exactly the same arguments can be used for the double ancestral.

One attractive way to argue for the ancestral and double ancestral as primitives is to argue that we grasp them by grasping the introduction and elimination rules for them, as with other logical vocabulary (helped by informal explication, again as with other logical vocabulary). This view of ancestral style predicates is urged by Parsons (2007, Chapter

[2]Others have also used ancestral logic as a middle ground between first and second order logic, such as Heck (2011, pp. 274–279), though Heck does not give sustained arguments for this status.

8), and exactly the same could be said of relations formed by the double ancestral.

We will see those these rules for the double ancestral in a second, but first we will characterize it in stricter logical terms. Formally the double ancestral is an operator on formulae that produces relation symbols. We introduce an extra clause into the recursive definition of formulae for the language: if ϕ and ψ are formulae, x_1, x_2 are distinct variables, y_1, y_2 are distinct variables, and s_1, s_2, t_2, t_2 are terms, then we obtain a formula $(\phi, \psi)^*_{x_1,x_2,y_1,y_2}(s_1, t_1, s_2, t_2)$. Free occurrences of x_1, x_2 in ϕ become bound in this formula, as do free occurrences of y_1, y_2 in ψ.

Now for the rules for the double ancestral. We will shortly see that it can be used to play the role of the (single) ancestral, so since there is no complete effective deductive system for the ancestral operator (Shapiro 1991) there isn't one for the double ancestral operator either. However we can still give natural deductive rules that capture the reasoning we use for it in practice. These parallel those for the ancestral described by Smith, and the rules for the predicate "natural number" described and defended by Parsons (2007, Chapter 8). It is these rules that will be used to test Isaacson's thesis, so that if one could argue that there were further inferences that a grasp of the double ancestral should license, the test of Isaacson's thesis would be undermined; there are no obvious candidates for this though. I use $\phi[t|x]$ to denote the substitution of the term t for free occurrences of variable x in ϕ.

$$\frac{s_1 = t_1 \qquad s_2 = t_2}{(\phi, \psi)^*_{\vec{x},\vec{y}}(s_1, t_1, s_2, t_2)} \tag{1}$$

$$\frac{(\phi, \psi)^*_{\vec{x},\vec{y}}(s_1, t_1, s_2, t_2) \qquad \phi[s_2|x_1, s_3|x_2] \qquad \psi[t_2|y_1, t_3|y_2]}{(\phi, \psi)^*_{\vec{x},\vec{y}}(s_1, t_1, s_3, t_3)} \tag{2}$$

$$\frac{\forall \vec{x}\vec{y}\,((\chi(x_1, y_1) \wedge \phi \wedge \psi) \Rightarrow \chi[x_2|x_1, y_2|y_1]) \qquad (\phi, \psi)^*_{\vec{x},\vec{y}}(s_1, t_1, s_2, t_2)}{\chi[s_1|x_1, t_1|y_1] \Rightarrow \chi[s_2|x_1, t_2|y_1]} \tag{3}$$

In rule (3) we require that x_2 and y_2 are not free in χ. The first two rules give ways of showing that objects lie under the double ancestral, the third is an induction rule: if

some property χ is preserved by ϕ together with ψ, then it is preserved by $(\phi, \psi)^*_{\vec{x},\vec{y}}$.

We now quickly sketch a semantics for this. If A is a structure for the language and v a variable assignment over A then we stipulate that $D, v \vDash (\phi, \psi)^*_{\vec{x},\vec{y}}(s_1, t_1, s_2, t_2)$ iff there exist sequences $(a_0, \ldots a_n)$ and $(b_0, \ldots b_n)$ for some $n \geqslant 0$ such that $a_1 = v(s_1)$, $b_1 = v(t_1)$, $a_n = v(s_2)$, $b_n = v(t_2)$, and for each $i = 0 \ldots (n-1)$ we have $A, v(x_1 \mapsto a_i, x_2 \mapsto a_{i+1}) \vDash \phi$ and $v(y_1 \mapsto b_i, y_2 \mapsto b_{i+1}) \vDash \psi$. Otherwise one can employ the double ancestral in the metalanguage for this clause.

Next we note that the double ancestral operator can be used to define the ancestral operator. If we have a relation $\phi(x, y)$ for which we wish to form the ancestral $(\phi)^*_{x,y}(w, z)$, we can take some variables u_1, u_2 distinct from x, y, w, z, take ψ to be "$u_1 = u_2$", and take $(\phi)^*_{x,y}(w, z)$ to be $\exists u_1((\phi, \psi)^*_{x,y,u_1,u_2}(w, u_1, z, u_1))$. It is an easy check that this has the right semantics and satisfies Smith's deductive rules.

The double ancestral defined here is a special case of the two place generalized ancestral, which was defined by R. M. Martin (1943). Avron (2003, pp.157–158) also discusses the generalized ancestral, and proves that it cannot be defined in terms of the ancestral. As we will soon see, the double ancestral can be used to define primitive recursive functions, so Avron's proof also shows that the double ancestral cannot be defined in terms of the ancestral. It is possible to define the generalized ancestral (and thus the double ancestral) in terms of the ancestral if one has a pairing function on objects, as we will see in proposition IV.1.

Thus one could conceivably try to sidestep the arguments made here by claiming that we grasp a pairing function for natural numbers, and thus that the double ancestral in an arithmetic context can be defined in terms of the ancestral, with us grasping primitive recursive functions via a pairing function combined with the ancestral in this way. However it seems clear that appealing to a pairing function for natural numbers would be a very bad explanation of our grasp of addition and multiplication. The ability to form a pair (m, n) of two natural numbers, either as a self standing object or via

an injection $\mathbb{N}^2 \to \mathbb{N}$, is no part of a usual understanding of PA. Pedagogically, our understanding of addition and multiplication has nothing to do with a pairing function $\mathbb{N}^2 \to \mathbb{N}$ – we learn addition and multiplication long before learning about a pairing function, and students are often surprised to discover that such an injection exists. Grasping abstract pairs or general tuples of natural numbers also seems to be no part of our initial conception of arithmetic. Thus it will be assumed that founding our grasp of primitive recursion on a pairing function is an unattractive option.

I focus on the double ancestral rather than the generalized ancestral in this chapter however because it allows a simpler informal characterization, and is a closer fit for the case of primitive recursive functions. One could argue that anyone who grasps the double ancestral should be able to grasp the two place generalized ancestral; whether or not that is correct, the conservativeness argument given later would also apply to the two place generalized ancestral, so Isaacson's thesis is safe either way.

4 Primitive recursion and the double ancestral

When Smith argues that a grasp of the ancestral is used to understand the predicate "natural number", he does so by pointing out that

> understanding quantification over the [natural] numbers involves understanding that the numbers are zero, the next number, the one after that, *and so on*, *without limit* – and understanding too that these are the *only* numbers. Which is in effect to grasp the thought that every number stands in the ancestral of the successor relation to zero. (Smith 2008, pp.3–4, emphasis his)

Fix an object a and a function f, and consider the primitive recursive function g

defined by

$$g(0) = a$$
$$g(S(n)) = f(g(n)).$$

a might be any object (not necessarily a number). Exactly parallel to the above explication of what it is to understand the predicate "natural number", we can say that

> understanding the function g involves understanding that g applied to zero gives a, that g applied to the next number after zero is f of g applied to zero, that g applied to the next number after that is f of g applied to that number, that g applied to the next number after that is f of g applied to *that* number, and so on.

Understanding some sort of informal explication along these lines is how we understand what we mean by g, and why we can introduce a function symbol with these properties – in exactly the same way as understanding the ancestral is how we know we can form the predicate "natural number". The above is doubtless less clear than the earlier explication of "natural number", but it has a parallel structure, just involving twice as many objects. It is also visibly an explication of the function g in terms of the double ancestral. This is illustrated diagrammatically in fig. IV.3.

Comparing this to fig. IV.1 and fig. IV.2 makes it pretty clear, I think, that definition by primitive recursive is a straightforward case of the double ancestral; and, in fact, that the double ancestral generalizes definition by primitive recursive in an exactly parallel way to how the ancestral generalizes the definition of the concept "natural number". We can see the same thing happening in prose. We can describe the above function g by saying $g(x) = y$ iff

- $x = 0$ and $y = a$

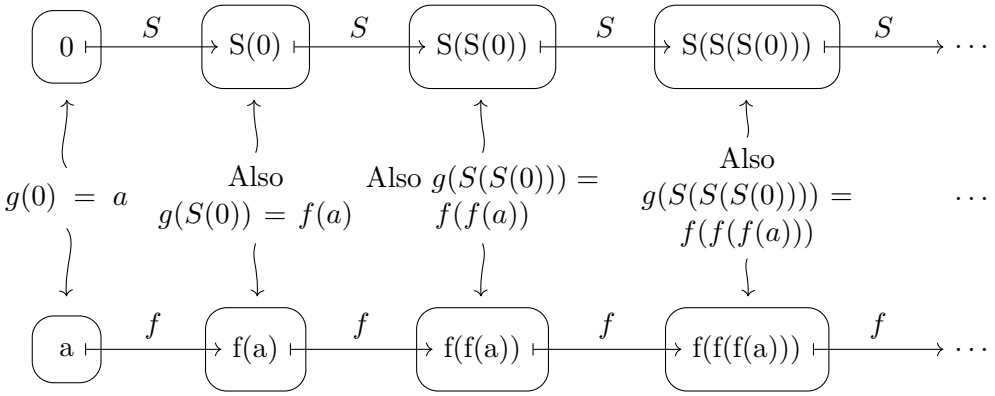


Figure IV.3: Primitive recursion via the double ancestral

- Or $x = S(0)$ and $y = f(a)$
- Or there are u and v such that $u = S(0)$ and $x = S(u)$, and $v = f(a)$ and $y = f(v)$
- Or there are u, u' and v, v' such that $u = S(0)$ and $u' = S(u)$ and $x = S(u')$, and $v = f(a)$ and $v' = f(v)$ and $y = f(v'')$
- Or there are u, u', u'' and v, v', v'' such that $u = S(0)$ and $u' = S(u)$ and $u'' = S(u')$ and $x = S(u'')$, and $v = f(a)$ and $v' = f(v)$ and $v'' = f(v'')$ and $y = f(v'')$

$$\vdots$$

and so on (and these are the only objects related by g). This is visibly an example of a definition of which the prose characterization of $(\phi, \psi)^*(c, d, x, y)$ seen in section IV.3 is the general form.

Exactly as Smith argues that grasping the concept natural number means grasping it as an instance of the ancestral operator, we can argue that grasping a definition by primitive recursion means grasping it as an instance of the double ancestral operator. There seems to be no good reason why anyone who can grasp a function defined by primitive recursion should not be able to grasp other instances of the double ancestral.

Formally, for this primitive recursive definition via the double ancestral to work all

we need is that $\forall x\ S(x) \neq 0$ and $\forall xy\ (S(x) = S(y) \to x = y)$, so that S is a suitable successor function, being injective and avoiding 0. Defining the ancestral $(\phi)^*_{x,y}(w, z)$ in terms of the double ancestral as seen in section IV.3 (or taking it as an extra primitive), we can then define a predicate $\mathbb{N}$ for the successors of 0 – the natural numbers – to apply to those x satisfying $(v = S(u))^*_{u,v}(0, x)$. We define the function g via the relation $G(x, y)$ defined as

$$(u_2 = S(u_1), v_2 = f(v_1))^*_{u_1,u_2,v_1,v_2}(0, a, x, y).$$

It is straightforward by induction along the predicate $\mathbb{N}$ that for all x satisfying $\mathbb{N}(x)$ there is a y such that $G(x, y)$, using the fact that f is total and rule (2) from section IV.3. Thus G is a total relation, and we seek to argue that it defines a well defined function. First, putting the relation $\chi(w, z)$ defined to be $w = 0 \to z = a$ into the induction rule for G, rule (3), gives that for all x, y, if $G(x, y)$ then $x = 0$ implies $y = a$; in other words, $G(0, y)$ implies $y = a$. Thus we have that $G(0, y)$ and $G(0, y')$ implies $y = y'$. Next we prove a lemma, arguing by induction along G that if $G(Sx, y)$ then there is z such that $y = g(z)$ and $G(x, z)$; the induction hypothesis is the relation $\chi(w, z)$ defined to be $G(w, z) \wedge (\forall u\ (w = S(u) \to \exists v\, (z = g(v) \wedge G(u, v))))$. Finally we argue by induction along the predicate $\mathbb{N}(x)$, that if x satisfies $\mathbb{N}(x)$ then for all y, y', if $G(x, y)$ and $G(x, y')$ then $y = y'$. The base case for 0 was the first thing proved above. For the induction step, if the property holds for x, and we have $G(Sx, y)$ and $G(Sx, y')$, then by the lemma there are z, z' such that $G(x, z)$ and $G(x, z')$ and $y = f(z)$, $y' = f(z')$. Thus by the induction hypothesis $z = z'$, and so $y = y'$ as required. Thus G does indeed define a total and well defined function, which we can denote as above by g. It is immediate by rules (1) and (2) that this function satisfies the defining equations mentioned initially in this section.

One possible caveat to the arguments of this section is that there are interpretations of arithmetic on which addition and multiplication might not be seen as given by pri-

mitive recursion: for instance the approach of Neo-Fregeanism via cardinality, and the interpretation of arithmetic in intuitive terms via strings of symbols given by Parsons (2007). As noted in section IV.1, a Neo-Fregean perspective does seem to present major problems for Isaacson's thesis, which I cannot address here. Parsons does argue that addition and multiplication in the context of strings should be seen as distinct from an intuitive perspective from other primitive recursive functions (ibid., Chapter 7), but this is very much a result about his particular notion of intuition, and this string based context. He does not argue that this string interpretation is the true interpretation of arithmetic, and does not argue that it gives an interpretation of all of PA. Thus his views (even if correct) do not present much of a challenge to the perspective here.

5 Double ancestral arithmetic

We call the theory of $\mathbb{N}$ mentioned in section IV.4 double ancestral arithmetic. It has a language with the constant 0 and the successor function S, and its axioms are:

$$\forall x\ (v = S(u))^*_{u,v}(0, x) \tag{4}$$

$$\forall x\, S(x) \neq 0 \tag{5}$$

$$\forall xy\, (S(x) = S(y) \rightarrow x = y). \tag{6}$$

This is a particularly simple and natural axiomatization – all we need is that every number is a successor of 0, and and that the successor function is injective without 0 in its range. It is categorical because of the standard semantics for the double ancestral, and thus for the relation $(v = S(u))^*_{u,v}$.

We can give an informal characterization of addition similar to that of general primitive recursive functions:

Adding 0 to x gives you x, adding 1 to x gives you Sx, adding 2 to x gives

you SSx, adding 3 to x gives you $SSSx$, and so on.

It follows that the relation $x + y = z$ can be captured by the double ancestral

$$(u_2 = S(u_1),\ v_2 = S(v_1))^*_{u_1,u_2,v_1,v_2}(0, x, y, z)$$

(one can visualise a diagram similar to fig. IV.3 to see how this works). As with the case of a general primitive recursive function seen in section IV.4, one can prove straightforwardly in double ancestral arithmetic that this definition does indeed define a total, single-valued function with value z of its arguments x and y. Using the normal notation $x + y = z$ for it, we have that $+$ satisfies the usual equations

$$x + 0 = x$$
$$x + Sy = S(x + y).$$

Similarly multiplication $x \times y = z$ is captured by the double ancestral

$$(u_2 = S(u_1),\ v_2 = v_1 + x)^*_{u_1,u_2,v_1,v_2}(0, 0, y, z).$$

Again one can prove its usual properties in the theory.

Thus one obtains the axioms of Smith's ancestral arithmetic and (*a fortiori*) of PA in double ancestral arithmetic – the instances of the induction axiom scheme follow from axiom (4). The double ancestral provides the general concept of which addition and multiplication are a special case. Thus it provides a useful test case for Isaacson's thesis.

If Smith is correct that a grasp of the ancestral is what's needed to grasp the predicate "natural number", and the parallel argument here for the double ancestral and primitive recursion is also correct, then the axiomatization of arithmetic above in terms of the double ancestral appears to include everything that is needed for a full understanding of PA (as long as the deductive rules for the ancestral and double ancestral are in some

sense adequate).

Thus conservativeness of double ancestral arithmetic over PA would imply that one would have to employ ideas beyond those needed to understand PA in order to prove a statement of L_A that was unprovable in PA. On the other hand if double ancestral arithmetic is not conservative over PA, then – on a very natural interpretation of arithmetic – we would have examples of statements of L_A provable using only the conceptual resources needed to understand PA, so Isaacson's thesis would be in trouble.

Fortunately for Isaacson's thesis, we have the following.

Proposition IV.1. *Double ancestral arithmetic is conservative over ancestral arithmetic (as defined by Smith).*

Proof. We show that ancestral arithmetic and double ancestral arithmetic are definitionally equivalent. Let T_{Anc} be the theory of ancestral arithmetic, L_{Anc} its language, and let T_{DA} be the theory of double ancestral arithmetic and L_{DA} its language. We saw above how to define the relevant primitives of ancestral arithmetic – the ancestral operator, addition and multiplication – in terms of the double ancestral. Let $\phi \mapsto f(\phi)$ denote the translation by these definitions from the L_{Anc} to L_{DA}. We have that if $T_{\mathrm{Anc}}, \Gamma \vdash$ then $T_{\mathrm{DA}}, f(\Gamma) \vdash f(\phi)$, since double ancestral arithmetic proves the axioms for the ancestral and addition and multiplication.

Now in ancestral arithmetic we can define a bijective pairing function $\alpha : \mathbb{N}^2 \to \mathbb{N}$, where $\alpha(x, y) = \alpha(x', y')$ iff $x = x'$ and $y = y'$. We let the inverse be $z \mapsto (\beta_1(z), \beta_2(z))$. We will show how to translate statements involving the double ancestral into statements involving the (standard) ancestral using this. This translation function will be denoted by g.

The idea is simply that $(\phi, \psi)^*_{x_1,x_2,y_1,y_2}(s_1, t_1, s_2, t_2)$ is equivalent to the ancestral

$$(\phi[\beta_1(x), \beta_1(y)] \wedge \psi[\beta_2(x), \beta_2(y)])^*_{x,y}(\alpha(s_1, t_1), \alpha(s_2, t_2))$$

This can be easily checked to be the case semantically.

We define g by induction on the number of occurrences of the double ancestral in a formula. For statements θ of L_{DA} which do not involve RTC^2, $g(\theta)$ is just θ. For a statement of the form $\theta = (\phi, \psi)^*_{x_1,x_2,y_1,y_2}(s_1, t_1, s_2, t_2)$, we let ϕ' be $g(\phi)[\beta_1(x)|x_1, \beta_1(y)|x_2]$ and ψ' be $g(\psi)[\beta_2(x)|y_1, \beta_2(y)|y_2]$, and then define $g(\theta)$ to be $(\phi' \wedge \psi')^*_{x,y}(\alpha(s_1, t_1), \alpha(s_2, t_2))$. g acts on statements built from propositional connectives or quantifiers in the obvious way, e.g. $g(\theta_1 \wedge \theta_2) = g(\theta_1) \wedge g(\theta_2)$. It is an easy check that this is an adequate definition of the double ancestral, in as much as the deductive rules for the double ancestral hold: we have that if $T_{\mathrm{DA}}, \Delta \vdash \theta$ then $T_{\mathrm{Anc}}, g(\Delta) \vdash g(\theta)$.

Then it is not difficult to show f and g give a definitional equivalence (Corcoran 1980), i.e. that for any $\chi \in L_{\mathrm{Anc}}$ we have $T_{\mathrm{Anc}} \vdash \chi \Leftrightarrow g(f(\chi))$, and for any $\theta \in L_{\mathrm{DA}}$ we have $T_{\mathrm{DA}} \vdash \theta \Leftrightarrow f(g(\theta))$.

Thus if χ is a formula of L_{Anc} and $T_{\mathrm{DA}} \vdash f(\chi)$ then $T_{\mathrm{Anc}} \vdash g(f(\chi))$ so $T_{\mathrm{Anc}} \vdash \chi$. In this sense T_{DA} is conservative over T_{Anc}, as claimed. □

Putting this together with Smith's result that ancestral arithmetic is conservative over PA, we can conclude that double ancestral arithmetic is conservative over PA. Thus Isaacson's thesis passes the test, and looks secure; even more secure than it did after passing Smith's test, since we have now taken the functions of addition and multiplication into account.

Incidentally this argument makes it clear that understanding primitive recursion does not require quantifying over relations, and is available on the basis of a purely logical, ontologically innocent operator.

Chapter V

Ancestrals and plurals

We now combine the idea of the ancestral and double ancestral with the resources of plural logic, briefly showing how this allows the definition of various key arithmetic concepts – finiteness, equinumerosity (for finite pluralities), addition and multiplication – and thus allows a novel kind of interpretation of arithmetic. Aspects of this are compared to the Neo-Fregean interpretation. If one accepts ancestral logic and plural logic separately as real logics, then there is a strong case that their combination here should also be considered a real logic, in which case we obtain that the concepts of finiteness and equinumerosity for finite pluralities, and the other arithmetic relations defined here, are purely logical. This will be useful for the issues considered in chapter VI.

Once again, the basic ideas here are in a sense not new, and follow the example of R. M. Martin (1943; 1949). He set up a nominalist system using a generalized ancestral operator together with a mereological base system, to provide a setting for arithmetic. The definitions of finiteness, equinumerosity and multiplication that I will give are essentially those given by him, except that I work with pluralities instead of the fusions of atoms (mereological simples) that he uses. His motivations were different: he appears to have been aiming just to give a fairly strong nominalistic system for doing maths in, and gave little philosophical discussion, not claiming his work had implications for

the semantics or epistemology of arithmetic. Nonetheless, all due credit to Martin. If nothing else, this chapter can be seen as advocacy of the importance of his work.

1 Plural double ancestral logic

The original case for plural logic was made by Boolos (1984; 1985), who argued that English contains statements involving plural quantification that cannot be translated away into purely singular terms (such as "there are critics who admire only one another"), and gave a semantics for plural logic, logic containing such plural quantifiers. A more extended case is made by Oliver and Smiley (2013), who argue that "[t]here are three good reasons to recognize plural terms: the notion is coherent; English itself appears to contain many examples; and rival singularist treatments of the phenomena fail" (ibid.). I think these arguments make a reasonable *prima facie* case for accepting plurals, though I do not think they settle the matter. For the purpose of this chapter and the following one, we will assume they are correct however.

Thus as well as singular variables x, y, $z, \ldots$ we introduce here plural variables xx, yy, $zz, \ldots$, with quantifiers over each type of variables, the usual logical rules, and comprehension for pluralities. It is slightly easier to define the arithmetic concepts discussed here if one allows an empty plurality, though this may be controversial, and the cases where pluralities are required to consist of at least one, or at least two, objects are also discussed.[1]

For plural variables there is an important distinction between collective and distributive predicates, discussed by Oliver and Smiley (ibid., §2.1): a distributive predicate states for instance that the men are rotund, where this applies to the men if and only

[1]The versions of plural logic in which "there are" means "there are zero or more", "there are one or more" or "there are two or more" respectively are interpretable in terms of each other, as Burgess and Rosen (1997, pp. 151, 155) discuss. It does not look like these reinterpretations work for the general case of ancestral and double ancestral operators operating on pluralities, but the reinterpretations may be valid for the instances of these operators used here.

if it applies to each individual, whereas a collective predicate states for instance that the men are carrying a stone, which can apply to the men even if it applies to none of them individually. Provided one accepts collective predicates, and similarly collective relations, for pluralities, and accepts the kind of rule based defence of the ancestral mentioned in section IV.3 urged by Parsons (2007, Chapter 8), then there appears to be no objection to accepting versions of the ancestral and double ancestral which form collective relations between pluralities; examples of such collective relations, such as the relation of equinumerosity, and of one plurality being a finite extension of a second, will be seen shortly.

Given a formula ϕ and distinct plural variables xx_1 and xx_2, the ancestral for pluralities gives a new binary relation symbol $(\phi)^*_{xx_1,xx_2}$ which takes two plural terms as arguments. We can informally explain $(\phi)^*_{xx_1,xx_2}$ in prose, paralleling the explanation of the double ancestral from section IV.3. We have that the relation $(\phi)^*_{xx_1,xx_2}(aa, tt)$ holds iff

- $aa = ss$
- Or $\phi[aa|xx_1, ss|xx_2]$
- Or there are uu such that $\phi[aa|xx_1, uu|xx_2]$ and $\phi[uu|xx_1, ss|xx_2]$
- Or there are uu and there are uu' such that $\phi[aa|xx_1, uu|xx_2]$ and $\phi[uu|xx_1, uu'|xx_2]$ and $\phi[uu'|xx_1, ss|xx_2]$
- Or there are uu and there are uu' and there are uu'' such that we have $\phi[aa|xx_1, uu|xx_2]$ and $\phi[uu|xx_1, uu'|xx_2]$ and $\phi[uu'|xx_1, uu''|xx_2]$ and $\phi[uu''|xx_1, ss|xx_2]$

$$\vdots$$

and so on (and these are the only pluralities related by $(\phi)^*_{xx_1,xx_2}$). This is the same as how one would informally explain the usual ancestral, but for plural variables instead.

Then the double ancestral creates a relation symbol $(\phi, \psi)^*_{xx_1,xx_2,yy_1,yy_2}$ which takes four plural terms as arguments, given formulae ϕ and ψ and distinct plural variables xx_1, xx_2, yy_1, yy_2. Again we can give an informal explication of this in prose, paralleling that of the double ancestral from section IV.3. We have that the relation $(\phi, \psi)^*_{xx_1,xx_2,yy_1,yy_2}(aa, bb, ss, tt)$ holds iff

- $aa = ss$ and $bb = tt$
- Or $\phi[aa|xx_1, ss|xx_2]$ and $\psi[bb|yy_1, tt|yy_2]$
- Or there are uu and there are vv such that $\phi[aa|xx_1, uu|xx_2]$ and $\phi[uu|xx_1, ss|xx_2]$, and $\phi[bb|yy_1, vv|yy_2]$ and $\phi[vv|yy_1, tt|yy_2]$
- Or there are uu and there are uu', and there are vv and there are vv', such that $\phi[aa|xx_1, uu|xx_2]$ and $\phi[uu|xx_1, uu'|xx_2]$ and $\phi[uu'|xx_1, ss|xx_2]$, and $\phi[bb|yy_1, vv|yy_2]$ and $\phi[vv|yy_1, vv'|yy_2]$ and $\phi[vv'|yy_1, tt|yy_2]$
- Or there are uu and there are uu' and there are uu'', and there are vv and there are vv' and there are vv'', such that $\phi[aa|xx_1, uu|xx_2]$ and $\phi[uu|xx_1, uu'|xx_2]$ and $\phi[uu'|xx_1, uu''|xx_2]$ and $\phi[uu''|xx_1, ss|xx_2]$, and $\phi[bb|yy_1, vv|yy_2]$ and $\phi[vv|yy_1, vv'|yy_2]$ and $\phi[vv'|yy_1, vv''|yy_2]$ and $\phi[vv''|yy_1, tt|yy_2]$

$$\vdots$$

and so on (and these are the only pluralities related by $(\phi, \psi)^*_{xx_1,xx_2,yy_1,yy_2}$).

These are governed by the same rules as for the ancestral and double ancestral, but for pluralities. Explicitly, we have

$$\frac{aa = ss}{(\phi)^*_{\vec{xx}}(aa, ss)} \tag{1}$$

$$\frac{(\phi)^*_{\vec{xx}}(aa, ss_1) \qquad \phi[ss_1|xx_1, ss_2|xx_2]}{(\phi)^*_{\vec{xx}}(aa, ss_2)} \tag{2}$$

$$\frac{\forall\, xx_1\, xx_2\ ((\chi(xx_1) \wedge \phi) \Rightarrow \chi[xx_2|xx_1]) \qquad (\phi)^*_{\vec{xx}}(aa, ss)}{\chi[aa|xx_1] \Rightarrow \chi[ss|xx_1]} \tag{3}$$

In rule (3) we require that xx_2 is not free in χ.

For the plural double ancestral we have

$$\frac{aa = ss \qquad bb = tt}{(\phi, \psi)^*_{\vec{xx},\vec{yy}}(aa, bb, ss, tt)} \tag{4}$$

$$\frac{(\phi, \psi)^*_{\vec{xx},\vec{yy}}(aa, bb, ss_1, tt_1) \qquad \phi[ss_1|xx_1, ss_2|xx_2] \qquad \psi[tt_1|yy_1, tt_2|yy_2]}{(\phi, \psi)^*_{\vec{xx},\vec{yy}}(aa, bb, ss_2, tt_2)} \tag{5}$$

$$\frac{\forall\, \vec{xx}\, \vec{yy}\ ((\chi(xx_1, yy_1) \wedge \phi) \Rightarrow \chi[xx_2|xx_1, yy_2|yy_1]) \qquad (\phi, \psi)^*_{\vec{xx},\vec{yy}}(aa, bb, ss, tt)}{\chi[aa|xx_1, bb|yy_1] \Rightarrow \chi[ss|xx_1, tt|yy_1]} \tag{6}$$

In rule (6) we require that xx_2 and yy_2 are not free in χ.

One can give a semantics for each of these operators using the same operator in the metalanguage, for instance extending the semantics for plural logic from Boolos (1985) appropriately.

2 Finiteness

To start characterizing arithmetic concepts, first we define what it is for a plurality yy to be the same as the xx but with one additional element, which we write as $\mathrm{Succ}(xx, yy)$, "Succ" for "successor". This holds if every x amongst the xx is also among the yy, and if there is a unique y which is amongst the yy but not the xx. Then using the ancestral for plurals we can form the relation $(\mathrm{Succ}(xx_1, xx_2))^*_{xx_1,xx_2}$, where $(\mathrm{Succ}(xx_1, xx_2))^*_{xx_1,xx_2}(ss, tt)$ holds if the tt are obtained as a finite extension of the ss. Then we can define what it is for a plurality to be finite: they are just the finite extensions of the empty plurality (or the finite extensions of the pluralities of size one or two, depending on what we take the minimum size of a plurality to be). The informal characterization of this is that the empty plurality is finite, and that if the uu are empty

and the ss are obtained from uu by the addition of a single element, then the ss are finite, and that if the uu are empty and the uu' are obtained from the uu by the addition of a single element, and the ss are obtained from the uu' by the addition of a single element, then the ss are finite, and so on. This is a natural informal characterization of finiteness.

Then the rules for the plural ancestral tell us that the empty plurality is finite (or all those of size one or two, depending on the minimum size of our pluralities), that if the ss are finite and the tt are the same as the ss but with one element added, then the tt are finite; and that if some property is preserved under addition of one element to a plurality, and holds of the empty plurality, then it holds of all finite pluralities. These principles seem very plausibly to be constitutive of what we mean by finiteness. The first two would I think be accepted by almost anyone; the inductive principle is a little more obscure, but I think we would expect people to accept at least some arguments along these lines. For instance, if we wanted to argue that every finite group of people has a tallest person, we could argue as follows: if there is just one among the xx then they have a tallest person, and if the yy have a tallest person u and people zz are just the yy with an extra person v then either u or v will be a tallest person among the yy. It is possible that many people would not see the need for this kind of argument (taking this kind of fact as obvious). It is also possible that an ordinary person might phrase it more naturally in a different way, as saying that if there's one person then they're the tallest; if you add a second person, then if they're taller than the previous person then they're the tallest, otherwise the previous person was the tallest; if you add a third person, then if they're taller than the previous two then they're the tallest, otherwise the tallest of the previous two is the tallest, and so on. Though phrased differently this has essentially the same structure as the preceding argument. It appears here that we can see an ordinary person's grasp of inductive inferences involving finiteness as on a similar sort of footing to for instance inferences involving $\forall$-introduction in first order

logic.

3 Equinumerosity

The definition of equinumerosity for finite pluralities is very similar, but using the plural double ancestral instead of the plural ancestral. Using the double plural ancestral we can form the relation $(\mathrm{Succ}(xx_1, xx_2), \mathrm{Succ}(yy_1, yy_2))_{x\vec{x},y\vec{y}}$, where

$$(\mathrm{Succ}(xx_1, xx_2), \mathrm{Succ}(yy_1, yy_2))_{x\vec{x},y\vec{y}}(aa, bb, ss, tt)$$

holding means that the ss are obtained from the aa by adding in as many additional elements as it takes to obtain the tt from the bb. We write this relation as $\mathrm{Add}(aa, bb, ss, tt)$. If we allow an empty plurality, one can then define ss and tt to be equinumerous if $\mathrm{Add}(ee, ee, ss, tt)$ where ee is the empty plurality. Otherwise if we require all pluralities to have size at least one (or two), then we can define ss and tt to be equinumerous if there are aa, bb of size one (or two) such that $\mathrm{Add}(aa, bb, ss, tt)$. The informal characterization of this is that either the ss and the tt are empty, or there are empty uu and vv such that $\mathrm{Succ}(uu, ss)$ and $\mathrm{Succ}(vv, tt)$, or there are empty uu and vv and there are uu' and there are vv' such that $\mathrm{Succ}(uu, uu')$ and $\mathrm{Succ}(vv, vv')$ and $\mathrm{Succ}(uu', ss)$ and $\mathrm{Succ}(vv', tt)$, and so on.

This a natural informal way to describe the concept of equal size: both pluralities are empty (or have size one, or two, depending on the minimum size of our pluralities), or both are obtained by adding one extra to an empty plurality, or both are obtained by adding one extra to pluralities related like that, or both are obtained by adding one extra to pluralities related like that, and so on. This is similar to a description of the pluralities on each side as built up by the addition one by one of the same number of elements – and similar too to a description in terms of counting each plurality and obtaining the same result.

The rules for the double ancestral then tell us that:

- The empty plurality is equinumerous to itself (or all those of size one or two are equinumerous)
- If ss and tt are equinumerous, and the ss' and the tt' are obtained by adding a single element to ss and tt respectively, then the ss' and the tt' are equinumerous
- If a relation R holding between pluralities ss and tt implies that it holds between ss' and tt' when ss' and tt' are obtained by adding a single element to ss and tt respectively, and also R holds between any empty pluralities, then R holds between all equinumerous pluralities

We will write this equinumerosity relation as $xx \approx yy$. This has the basic properties one would expect. Indeed it is easy to see that if $xx \approx yy$ then xx and yy are finite, and that if xx is finite then $xx \approx xx$. Then the $\approx$-relation is symmetric for empty pluralities, and if $yy \approx xx$ and $\mathrm{Succ}(yy, yy')$ and $\mathrm{Succ}(xx, xx')$ then $yy' \approx xx'$, so it is immediate by $\approx$-induction on $xx \approx yy$ (with induction hypothesis $yy \approx xx$) that $xx \approx yy$ implies $yy \approx xx$. Also, by $\approx$-induction we have that if xx is not empty and $xx \approx yy$ then there are xx' and yy' such that $xx' \approx yy'$, $\mathrm{Succ}(xx', xx)$ and $\mathrm{Succ}(yy', yy)$ (call this the successor equinumerosity property). Thus if yy is empty and $xx \approx yy$ then xx must be empty.

Arguing that the relation is transitive appears to require more work. First, say that yy and yy' *differ in one element* if there is a unique y amongst the yy but not the yy', and a unique y' amongst the yy' but not amongst the yy. We can argue by induction on yy that if yy and yy' differ in one element then $yy \approx yy'$. This is trivial for yy empty. Suppose it holds for yy, and that we have $\mathrm{Succ}(yy, zz)$, with a unique x amongst the zz but not the yy; and that we have that zz' differs from zz in one element, with a unique z amongst the zz but not the zz' and a unique z' amongst the zz' but not the zz. If $x \neq z$ then we let yy' consist of the same elements as yy but with z' added and z removed; thus

yy' differs from yy in one element, so by the induction hypothesis we have $yy \approx yy'$, but we have $\text{Succ}(yy, zz)$ and $\text{Succ}(yy', zz')$ by construction, and so $yy' \approx zz'$ as required. If on the other hand $x = z$ then $\text{Succ}(yy, zz')$ and so $zz \approx zz'$. Thus either way we are done. Then we have that if $\text{Succ}(yy, zz)$ and $\text{Succ}(yy', zz)$ then either $yy = yy'$ or the yy differ from the yy' in one element, and so either way $yy \approx yy'$. Now we argue by induction on xx that if $xx \approx yy$ and $yy \approx zz$ then $xx \approx zz$. This is true for xx empty. Suppose true for xx, and that $\text{Succ}(xx, xx^*)$, and let $xx^* \approx yy$ and $yy \approx zz$. By successor equinumerosity, there are xx', yy', yy'', and zz' such that $xx' \approx yy'$, $\text{Succ}(xx', xx^*)$, $\text{Succ}(yy', yy)$, $yy'' \approx zz$, $\text{Succ}(yy'', yy)$ and $\text{Succ}(zz', zz)$. Then we also have $xx \approx xx'$ and $yy' \approx yy''$ by the first intermediate result. Thus $xx \approx xx' \approx yy' \approx yy'' \approx zz'$, and so by the induction hypothesis $xx \approx zz'$. Thus $xx^* \approx zz$ as required. Thus by induction whenever $xx \approx yy$ and $yy \approx zz$ we have $xx \approx zz$, so that $\approx$ is transitive, and is thus an equivalence relation on finite pluralities.

We can then also argue that if $\text{Succ}(xx, xx^*)$ and $\text{Succ}(yy, yy^*)$ then $xx \approx yy$ iff $xx^* \approx yy^*$. The only if direction is immediate from the rules governing $\approx$, and for the converse, we have by the second intermediate result proved on the way to transitivity that if $xx^* \approx yy^*$ then there are xx' and yy' with $xx' \approx yy'$, $\text{Succ}(xx', xx^*)$ and $\text{Succ}(yy', yy^*)$. Thus $xx \approx xx'$, and $yy \approx yy'$, so since $\approx$ is an equivalence relation we have $xx \approx yy$ as required.

This account of equinumerosity is an attractive one. As Burgess (2005, pp.80–85) discusses, when giving an account of a concept we can either intend a *hermeneutic* account, which is intended to describe what the term presently means, or a *revolutionary* account, which is purely a proposal for the future use of the concept. Giving a hermeneutic account is potentially a much greater challenge, as one needs to take into account what current or prior users of the concept have taken it to mean, whereas the primary question for a revolutionary account of a concept is just whether it is useful. The main existing account of equinumerosity is found in Hume's principle, stating that the xx are

equinumerous to the yy iff there is a relation R which is a bijection between the xx and the yy. This is essentially the definition of equinumerosity used in mathematics,[2] to which the account given here is not intended as a rival: in mathematics one wants powerful, mathematically fruitful concepts, and the approach used here appears to just be more awkward and less widely applicable than the definition in terms of bijections (as it does not apply to infinite sets).

However there is also the everyday, preexisting notion of equinumerosity, which needs separate consideration. As a revolutionary proposal for an everyday language notion of equinumerosity, the concept described above has the advantage that it is phrased in (arguably) a more innocent logic than Hume's principle, which requires second order logic: both the ancestral operator and the use of plurals have been argued to be purely logical, and ontologically innocent, and if this is right then the same should apply to the combination of them discussed here. The status of full second order logic by contrast is more questionable, relying as it does on the existence of abstract objects such as relations. As a hermeneutic account of the everyday concept of equinumerosity, the proposal here also has advantages. It has been noted, for instance by Heck (2011, pp.168–172), that a grasp of equinumerosity in terms of bijections does not seem to be essential to an everyday understanding of the concept. Someone can understand for instance the statement that "there will come a time as many days in the future as there are birds in the sky", with no awareness that this need involve a way of pairing up days and birds, either as a relation definable in our language, or as an abstract relation entity. If our pre theoretical notion of equinumerosity did involve a tacit understanding of bijections, then the extension to the infinite case ought to be straightforward for people, whereas the definition of cardinality in terms of bijections for infinite sets is an aspect of higher mathematics that is famously found to be difficult and unintuitive. By contrast, the three principles seen above that characterize equinumerosity on this account are much more natural. The

[2]Though in mathematics one takes a bijection to be a set of ordered pairs, instead of an entity in the range of a second order binary relation variable.

first two – that the empty plurality is equinumerous to itself (or all of size one or two are equinumerous, depending on our stance), and that if ss and tt are equinumerous, and the ss' and the tt' are obtained by adding a single element to ss and tt respectively, then the ss' and the tt' are equinumerous – are clearly basic to our standard conception of equinumerosity. These first two principles are actually very similar to two of the three principles that Heck takes to be basic to the concept Heck (ibid., p. 170): the second above is essentially the same principle as Heck's second principle (though phrased in plural rather than second order logic), whilst Heck's first principle is the first of the principles above conjoined with the claim that if xx is empty and is equinumerous with yy, then yy is empty, which is a property that we can easily derive, as noted above. Heck's third principle states essentially that if $S(xx, xx^*)$ and $S(yy, yy^*)$ and $xx^* \approx yy^*$ then $xx \approx yy$, which was also derived without too much difficulty above. The third principle governing equinumerosity here, the induction principle, is more difficult, but we can argue that at least some instances of it would naturally be accepted. As an example, we can argue by induction that whenever there are as many people xx_1 as people xx_2 then the xx_1 and the xx_2 can be lined up facing each other, with each person in each line directly opposite one person in the other line. This is true whenever there is a single xx_1 and xx_2. If it's true for xx_1 and xx_2 and the yy_1 are the xx_1 with extra person v_1, and the yy_2 are the xx_2 with extra person v_2, then xx_1 and xx_2 can be lined up facing each other, with v_1 and v_2 facing each other at one end. One could give a more informal version of this, but with a similar argumentative structure, by arguing that the conclusion held when there was one among the xx_1 and one among the xx_2, or two among the xx_1 and two among the xx_2, and so on, adding in an extra person to the argument at each step. It appears to be about as reasonable to take acceptance of this kind of inference as part of a standard notion of equinumerosity as it is to take a grasp of $\forall$-introduction to be implicit in a standard understanding of quantification. As other points in favour of the characterization given here, it avoids using the concept

of bijection, and has no obvious extension to the infinite case – fitting the layman's confusion about what "sameness of size" might mean for infinite totalities.

Though this conception does not employ the notion of bijection, one can still prove intuitive facts about bijections, such as that bijections between finite pluralities establish equinumerosity between them: if $\phi(x, y)$ is an open formula which is bijective in its two arguments, then we can prove by induction on xx that if ϕ relates the objects amongst the xx to the objects amongst the yy then $xx \approx yy$. Principles like this appear to underlie examples like Heck's of the cookies and the children (Heck 2011, p. 171).

4 Arithmetic operations

We now move on to discuss further arithmetic concepts, namely how to characterize what it is for a finite plurality zz to be the same size as as the sizes of the finite pluralities xx and yy added, or multiplied. For addition, we use the same relation Add as that in the definition of equinumerosity, and define $+(xx, yy, zz)$ to mean (assuming we allow the empty plurality) that there is a plurality xx_2 equinumerous with xx such that $\mathrm{Add}(ee, xx_2, yy, zz)$ where ee is the empty plurality (one can adjust this appropriately if one takes pluralities to be nonempty). This states informally that zz can be obtained from zz' by adding as many elements as it takes to obtain yy from ee, i.e. as many elements as there are in yy.

We can again prove basic properties of this that one would expect. It is clear from the definition that $xx \approx xx'$ and $+(xx, yy, zz)$ implies $+(xx', yy, zz)$. From the induction rule for the relation Add we obtain that if $\mathrm{Add}(ee, xx_2, yy, zz)$ with yy nonempty then there are yy' and zz' with $\mathrm{Succ}(yy', yy)$ and $\mathrm{Succ}(zz', zz)$ and $\mathrm{Add}(ee, xx_2, yy', zz')$. Thus if we have $+(xx, yy, zz)$ then there are yy' and zz' with $\mathrm{Succ}(yy', yy)$ and $\mathrm{Succ}(zz', zz)$ and $+(xx, yy', zz')$. It then follows without much difficulty by induction on yy that if $yy \approx yy_2$ then $+(xx, yy_2, zz)$ implies $+(xx, yy, zz)$. The base case where yy is empty

is trivial. For the induction step, we suppose the conclusion holds for yy, and that we have $\mathrm{Succ}(yy, yy^*)$ and $yy^* \approx yy_2^*$ with $+(xx, yy_2^*, zz)$. Then as noted there are yy_2 and zz' with $\mathrm{Succ}(yy_2, yy_2^*)$ and $\mathrm{Succ}(zz', zz)$ and $+(xx, yy_2, zz')$, so that $yy \approx yy_2$ and thus by the induction hypothesis $+(xx, yy, zz')$ and so $+(xx, yy^*, zz)$ as required. Thus if $xx \approx xx'$ and $yy \approx yy'$ then $+(xx, yy, zz)$ iff $+(xx', yy', zz)$.

Now we argue that if $zz \approx zz_2$ and $+(xx, yy, zz)$ implies $+(xx, yy, zz_2)$. We do this by induction on yy. For the base case, using the induction rule for Add we obtain that if yy is the empty plurality then $\mathrm{Add}(ee, xx_2, yy, zz)$ implies $xx_2 = zz$, so that $+(xx, yy, zz)$ implies $xx \approx zz$. Conversely of course if yy is the empty plurality then $xx \approx zz$ implies $+(xx, yy, zz)$. Thus if yy is the empty plurality then $+(xx, yy, zz)$ iff $xx \approx zz$, from which the base case of our induction follows immediately. For the induction step, we suppose we have a plurality yy for which the induction hypothesis holds, and that $\mathrm{Succ}(yy, yy^*)$, with pluralities zz, zz_2 such that $zz \approx zz_2$ and $+(xx, yy^*, zz)$. Then as noted in the previous paragraph there are yy' and zz' such that $\mathrm{Succ}(yy', yy^*)$ and $\mathrm{Succ}(zz', zz)$ and $+(xx, yy', zz')$. Thus $yy' \approx yy$, so that by the result of the previous paragraph we have $+(xx, yy, zz')$. Then also if we let zz_2' be the same as zz_2 but with one element removed then $\mathrm{Succ}(zz_2', zz_2)$, and so $zz_2' \approx zz'$, and thus by the induction hypothesis $+(xx, yy, zz_2')$, and thus $+(xx, yy^*, zz_2)$ as required. Thus if $xx \approx xx'$ and $yy \approx yy'$ and $zz \approx zz'$ then $+(xx, yy, zz)$ iff $+(xx', yy', zz')$. In effect, if we think of cardinalities as given by quotienting finite pluralities by the equinumerosity relation, then this says that the addition relation is a well defined relation on cardinalities. It is then easy to also show by induction on yy that $+(xx, yy, zz)$ and $+(xx, yy, zz_2)$ implies $zz \approx zz_2$, so that addition is a well defined (possibly partial) operation on cardinalities.

As noted in the previous paragraph, if ee is the empty plurality then we have $+(xx, ee, zz)$ iff $xx \approx zz$. Also, if $\mathrm{Succ}(yy, yy^*)$ and $\mathrm{Succ}(zz, zz^*)$ then $+(xx, yy, zz)$ implies $+(xx, yy^*, zz^*)$. These principles parallel the equations for addition in Peano Arithmetic.

Finally, the case of multiplication. We define $\times(xx, yy, zz)$ to be the relation

$$(\mathrm{Succ}(uu_1, uu_2), +(vv_1, xx, vv_2))_{\vec{uu},\vec{vv}}(ee, ee, yy, zz)$$

whose informal meaning is that zz is obtained from ee by repeatedly adding adding as many elements as there are in xx to ee, as many times as there are elements in yy. It is immediate then, by our results about addition, that if $xx \approx xx'$ then $\times(xx, yy, zz)$ iff $\times(xx', yy, zz)$. The induction rule for the relation $\times$ gives that if $\times(xx, yy, zz)$ with yy nonempty then there is yy' with $\mathrm{Succ}(yy', yy)$ and zz' with $+(zz', xx, zz)$ such that $\times(xx, yy', zz')$. It then follows as before by induction on yy that if $yy \approx yy_2$ and $\times(xx, yy_2, zz)$ then $\times(xx, yy, zz)$. Then, similarly to for addition, one obtains that if ee is the empty plurality then $\times(xx, ee, zz)$ iff $zz = ee$. Thus we obtain by induction on yy that if $\times(xx, yy, zz)$ and $zz \approx zz_2$ then $\times(xx, yy, zz_2)$, with the induction step following by a similar (and in fact simpler) argument to that for the corresponding result in the case of addition. Thus if $xx \approx xx'$, $yy \approx yy'$ and $zz \approx zz'$ then $\times(xx, yy, zz)$ iff $\times(xx', yy', zz')$. We also obtain easily, as before, by induction on yy that if $\times(xx, yy, zz)$ and $\times(xx, yy, zz_2)$ then $zz \approx zz_2$. Finally we have seen that if ee is the empty plurality then $\times(xx, ee, zz)$ iff $zz = ee$, and it is immediate that if $\mathrm{Succ}(yy, yy^*)$ and $+(zz, xx, zz^*)$ then $\times(xx, yy, zz)$ implies $\times(xx, yy^*, zz^*)$, again obtaining the analogues of the Peano axioms for multiplication.

Though these operations of successor, addition and multiplication parallel the familiar ones, there is no guarantee yet that they are total: if xx and yy are pluralities, there need be no plurality zz such that $S(xx, zz)$, $+(xx, yy, zz)$ or $\times(xx, yy, zz)$. To ensure that these operations are total, we need that there be infinitely many objects. In this context we have a very simple axiom of infinity: that there is a plurality yy which is not finite. It is immediate by induction on xx that if xx is finite then any subplurality of xx is finite, so that the axiom of infinity is equivalent to the statement that the plurality of

all objects is infinite. Thus given the axiom of infinity, no finite plurality contains every object that there is, and so for every finite plurality xx there is x' with x' not amongst xx, and thus if we add x' to the xx to obtain xx^* then we have a plurality satisfying $S(xx, xx^*)$. It follows by induction on yy that if xx and yy are finite then there is zz such that $+(xx, yy, zz)$, and then by induction on yy that if xx and yy are finite then there is zz such that $\times(xx, yy, zz)$. Thus successor, addition and multiplication are in effect total functions.

5 Abstraction and cardinalities

As has been mentioned above, it is natural on this approach to think of finite cardinalities as obtained by quotienting our finite pluralities by the equinumerosity relation. We can make this precise by introducing an abstraction principle

$$N(xx) = N(yy) \leftrightarrow xx \approx yy$$

in which we attach a cardinality, $N(xx)$, to each finite plurality xx – we need to restrict to finite pluralities as if xx is infinite then $xx \not\approx xx$ so the abstraction principle would require $N(xx) \neq N(xx)$.[3] N functions here as a term forming operator. We can call this principle Martin's principle, after R. M. Martin (1943; 1949), whose definition of equinumerosity was very similar to that given here, though he was working with fusions of atoms instead of with pluralities. There are two importantly different ways of understanding this abstraction principle, depending on whether we take the terms N forms to have the same type as our other terms, or to be of their own, separate type. The former is the impredicative option, the latter the predicative option. As Linnebo (2018) argues, predicative abstraction principles are much more innocent than impredicative

[3]One would thus have to either restrict the range of the plural variables to finite pluralities – or perhaps have multiple different types of plural variables – or bring in a free logic in which the term $N(xx)$ fails to denote if xx is not finite.

abstraction principles, giving truth conditions for statements involving the operator N in terms of statements in which N does appear; this is not possible in general for impredicative abstraction principles, which function more like a kind of implicit definition, not totally dissimilar to just taking the arithmetic vocabulary to be implicitly defined by its second order axiomatization, for instance. As with other kinds of implicit definition, things can go wrong with impredicative abstraction principles, with some leading to inconsistent theories. This is not the case for predicative abstraction principles. Thus the predicative version of the principle is the one I would advocate, and is the one that will be used here (though as will be noted shortly, one could also use the impredicative version as an attractive replacement for Hume's principle to obtain a new version of Neo-Fregeanism).

With the abstraction principle in place, we obtain relations S, $+$, $\times$ on finite cardinalities n, m, p paralleling the relations S, $+$ and $\times$ for pluralities, where $S(m, n)$ holds iff there are xx, yy such that $m = N(xx)$, $n = N(yy)$ and $S(xx, yy)$, $+(m, n, p)$ holds iff there are xx, yy and zz such that $m = N(xx)$, $n = N(yy)$, $p = N(zz)$ and $+(xx, yy, zz)$, and $\times(m, n, p)$ holds iff there are xx, yy and zz such that $m = N(xx)$, $n = N(yy)$, $p = N(zz)$ and $\times(xx, yy, zz)$. It follows from the above facts about these relations for pluralities that $S(m, n)$ and $S(m, n')$ implies $n = n'$, so that S gives a partial function on finite cardinalities. Similarly $+(m, n, p)$ and $+(m, n, p')$ implies $p = p'$ and $\times(m, n, p)$ and $\times(m, n, p')$ implies $p = p'$, so that $+$ and $\times$ are partial binary operations on finite cardinalities. For these operations to be total, we need the above mentioned axiom of infinity – totality of these operations for finite cardinalities then following easily from the discussion there.

Thus though predicative abstraction principles have a degree of innocence that impredicative abstraction principles lack, our choice of the predicative version of Martin's principle means that we only obtain a full theory of arithmetic if the world contains infinitely many non arithmetic objects. This seems to me to be a reasonable outcome –

that natural numbers are the cardinalities of finite pluralities, so that if there are only finitely many things then there are only finitely many natural numbers. When I have informally discussed these issues with non specialists, a view of arithmetic something along these lines – where numbers are the sizes of finite collections of objects – is often suggested, and when it is mentioned that the universe might only contain finitely many objects, so that there would then only be finitely many numbers, this is often taken to be a reasonable – though surprising – consequence (in my experience of the matter). On the resulting view of arithmetic, our knowledge of arithmetic may be partly logical – knowledge of plural double ancestral logic – and partly empirical, with it being an empirically established fact that there are infinitely many objects (if there are, in fact, infinitely many objects). This seems to me to be a reasonable result.

If one objects to this conclusion then one could instead use the impredicative version of Martin's principle, obtaining a version of Neo-Fregeanism in which equinumerosity is given by the above definition in plural double ancestral logic instead of by the definition in terms of bijections. As discussed above, this definition of equinumerosity seems to be a more attractive one, better fitting with intuitive ideas about equinumerosity, and avoiding the problems with the bijection version – such as that it suggest the infinite case should be intuitive, when people find it not to be. If one takes the impredicative version of Martin's principle then one does obtain that there are infinitely many finite cardinalities. The key fact in this argument is that (using set formation notation for pluralities) $\{N(xx) \mid xx \subsetneq yy\} \approx yy$, which isn't hard to show by induction for yy finite. Then it follows that $S(yy, \{N(xx) \mid xx \subseteq yy)\})$ (showing along the way that if $xx \subsetneq yy$ then $\neg(xx \approx yy)$), and thus that every natural number has a successor.

Chapter VI

Sound Foundations

1 Introduction

What do we want from a foundation for mathematics? The original foundational project of establishing a basis for mathematics maximally impervious to doubt is long dead, as Shapiro (1991) recounts (securing the coffin with extra nails in the process). The question of what a foundation can or should amount to is now contentious.

Some authors still perceive a substantial role for a foundation, with Linnebo and Pettigrew (2011) regarding a foundation as being an account of a particular part of reality, containing claims about its subject matter and backed up by a justification of those claims – (implicitly) thus providing a subject matter for and justification of the mathematics that can be carried out in it. Ladyman and Presnell (2018) go further, advocating a view on which a foundation consists of five components: a formal framework for mathematics, a semantics stating how the terms and rules of this framework are to be understood, a metaphysical underpinning for this semantics, an epistemology for these, and a methodology for mathematical practice.

Others are less demanding. Tait (2005) and Muller (2004) argue – in different ways – that the key feature of an axiom system is just its consistency, and that any consis-

tent axiom system implicitly defines a concept which then serves as its subject matter, appealing to a Wittgensteinian conception of meaning as use (where the axioms serve to determine the use of the concept, and thus its meaning). This can be seen as a much more lightweight view of what we require from a foundation, though neither uses the term "foundation" as such.

Maddy (2011) picks out a middle course, in a sense. She focuses more explicitly on the practice of mathematics, discussing various important mathematical roles set theory plays as a foundation – such as ruling on questions of coherence and existence in mathematics, facilitating interconnections between disparate branches of mathematics, and answering questions of provability and refutability (ibid., p. 34). She regards consistency as a basic requirement of an axiom system (ibid., p. 73), but on her view the key characteristic that should be used to decide between different choices of axioms is "mathematical depth", also described as mathematical fruitfulness, mathematical effectiveness and so on (ibid., pp. 80–83 100, 112, 116–177). Though she is not concerned with interpreting or intrinsically justifying the axioms – in fact, she concludes by arguing that intrinsic justifications are less valuable than extrinsic ones (ibid., §V.4) – she intends for the notion of mathematical depth to give a reasonably objective sense in which one axiom system can be preferable to another.

Finally, Awodey (1996; 2004) opposes the very idea of a foundation, advocating instead a "categorical-structural" approach in which every piece of mathematics is accompanied by a specification of the rules and axioms that it in particular uses (Awodey 2004, p. 56). The conclusions of mathematical arguments are then taken to be schematic statements, rather than as being true of a fixed domain of quantification (ibid., p. 59). This approach is intended to be truer to the actual practice and aims of mathematics. He believes that the standard foundational goal, of elaborating a particular system which provides enough objects for all the usual needs of mathematics, and enough rules and axioms for all the usual ways of reasoning about them, is misguided (ibid., pp. 55–56).

Awodey's position is supported by McLarty (2012), who states that it should be taken seriously.

These views exhibit a great range of thought concerning the desired nature and role of a foundation for mathematics. The positions of Linnebo and Pettigrew (2011) and Ladyman and Presnell (2018) are closest to what has traditionally been thought of as the role of a foundation, but they must face the worry that the questions of interpretation and justification they raise are of interest only to philosophers, and irrelevant to the actual business of mathematics. Certainly the question of whether the axioms of set theory are true is not one that many mathematicians express much interest in. The views of Maddy (2011) and Awodey (1996; 2004) by contrast are much more directly motivated by considerations internal to mathematics, and the focus of Tait (2005) and Muller (2004) on the consistency of axiom systems can also be seen as reflecting the most oft stated concern of mathematicians.

This chapter argues for a new perspective on the purpose and function of a foundation for mathematics, and on what the desiderata for deciding on a foundation should be. The aim is to argue for this perspective from concerns that are of importance to mathematics, rather than being purely philosophical; but one conclusion will be that the best way to champion a particular foundation appears be to defend an interpretation on which its axioms are true, thus leading us back to views like that of Linnebo and Pettigrew (2011) and Ladyman and Presnell (2018).

We start in section VI.2 by discussing the notion of proof in mathematics (again), noting that there are different senses of the term. We can talk about "proof in S" where S is any proof system – a system of formal rules determining which strings of symbols do or do not count as a formal proof. Though such a system only directly determines a notion of formal proof, it can also give rise to a notion of higher level informal proof, as discussed for the case of ZFC in chapter I; and by "proof in S" we may mean either the formal or informal notion. Either way, when we talk of proof in this sense it need not

have any epistemic connotations. We can talk about proof in S for any S, no matter how implausible its principles; and one can even work in a proof system with inconsistent axioms, in which case one may be able to prove anything, which of course tells us nothing about how likely any such conclusion is.

Then we also have the notion of genuine proof, or proof in the epistemic sense: a proof that establishes its conclusion as true. Here we talk simply of proving ϕ, without mentioning a system S, though in rigorous mathematics this will be proof according to the principles of some formal system. On the view put forward here, the key property of genuine proof is the following:

1. Genuine proof is factive: for p to be a genuine proof of ϕ, ϕ must in fact be true.

(This is the first of three key claims made in this chapter). One challenge to this account of genuine proof would state that a genuine proof is required merely to show that *if* the axioms are true, then the conclusion follows (similar to the view of Awodey (2004) mentioned above); I consider this if-thenist view, and argue that it is not in conflict with genuine proof being factive, but in fact is motivated by this very constraint.

Since a genuine proof establishes its conclusion as true, we need some interpretation of what the conclusion of such a proof means, and when it is true or false. Section VI.3 discusses various interpretations of mathematical statements that have been put forward. Not all such interpretations are acceptable, and to prepare the ground for the discussion of which are, section VI.4 discusses the idea that we can potentially obtain realizations even of abstract formal structural concepts. Then section VI.5 argues for a condition that interpretations should satisfy. Namely for a given interpretation of mathematics to be acceptable, it needs to be the case that:

2. If a mathematical generalization ϕ about a kind of structure is true (under the interpretation), it must be the case that ϕ actually holds of all realizations of that kind of structure.

This I term the eliminative constraint. The exact meaning of this constraint will be clarified.

With this all in hand, we are in a position to discuss what the purpose of a designating a proof system as a "foundation for mathematics" is. In section VI.6 we will see that candidate foundations are proof systems in which a wide range of mathematical structures can be represented and investigated, and it will be argued that:

3. The purpose of choosing a foundation for mathematics is to provide a system whose proofs of conclusions about as many different kinds of structures as possible – the natural numbers, the real numbers, groups, rings, topological spaces, banach spaces, manifolds, schemes, and so on – are then regarded as genuine proofs.

Though I think these three points are of independent interest, as a consequence of them we obtain the main conclusion of the chapter:

- When judging a candidate foundations S for mathematics, a key desideratum is that for as many kinds of structure as possible, whenever we prove in S a mathematical generalization ϕ about that kind of structure, ϕ actually holds of all realizations of that kind of structure.

This condition holding for a given kind of structure is the condition of soundness for S with regard to that kind of structure. That this should hold for as many kinds of structure as possible is advocated here as the primary condition we should require of a foundation for mathematics. For most kinds of structure this is a stronger condition than consistency, and much harder to argue for by just exploring the consequences of a system.

Indeed in section VI.7 it will be argued that the only plausible route to the soundness of a general proof system involves arguing for the soundness of set theory by giving some conception underlying the axioms on which they can be established to be true, such as the iterative conception of set, or the limitation of size conception: if we believe some such

conception, and believe that this justifies the axioms of ZFC, then we have a powerful way to argue for the soundness of ZFC. Moreover, this kind of argument appears to be the only live option for arguing for soundness in general. The soundness of ETCS and homotopy type theory may then be justifiable by proving a relative soundness result, giving an interpretation of them in terms of ZFC. The implications of this for the above views of foundations, and the views of Maddy (2011) on set theory and its interpretation, are discussed in section VI.9.

2 Proof, and proof

We start by considering again the concept of proof in mathematics. As mentioned in section VI.1, it will be argued that there are different senses of the term.

Any formal proof system T comes equipped with a precise notion of what it is for a string of symbols to be a formal proof (or formal derivation) in T. Then provided this notion is sufficiently workable for humans, we can obtain from it a notion of informal, high level proof in T, as discussed for the case of ZFC in chapter I: we can start by getting used to writing very detailed natural language arguments that are obviously formalizable in T, as with the week 2 level of detail from sections I.3 and I.4, and then gradually ascend to more and more compressed arguments, where inferences at each new level of compression are provable at a level of greater detail that we have already got comfortable with. In this way proofs at every stage can still (in principle) be written out as fully formal proofs in T, as discussed in section I.7.

This process can potentially give us a notion of informal proof in T for a range of systems T. The familiar case is ZFC. Then it is often claimed by proponents of categorical set theory (ETCS) that this can be viewed as an underlying framework for much of mathematics in much the same way that ZFC can, so that sufficiently high level informal proof in ETCS is very similar or indistinguishable to high level informal

proof in ZFC.[1] One of the stated goals of the homotopy type theory (HoTT) book was to develop a new style of "informal type theory" (Univalent Foundations Program 2013, p. iv), or informal proof in HoTT.

Informal proof in an arbitrary system need not provide a justification for its conclusion, or require an interpretation of it. One can for instance talk perfectly happily of proving statements of set theory or arithmetic in Quine's system New Foundations, despite its axioms lacking much in the way of intrinsic motivation; and one can even talk of proving statements in an inconsistent system, with anything then being provable (under classical logic) – this of course providing no warrant for believing the conclusions reached at all. Thus instead of talking about "proof in T" it might be more appropriate to talk about derivations in T, where it is understood that these can be either formal or informal – but either way, need have no semantic content, nor give any justification for their conclusion.

This notion of "proof/derivation in T" is thus to be contrasted with another notion of proof in mathematics, where we talk simply of proving ϕ, without mentioning a proof system T. This could be called the notion of genuine proof, or proof in the epistemic sense. Though not explicitly phrased in terms of a proof system, in rigorous mathematics this will always be a derivation in T, for some proof system T, as will be discussed in section VI.6.

The key property of this notion of proof is that it is factive – such a proof establishes its conclusion as true. Suppose for instance that a young mathematician comes forward with a shocking claim: they have a proof that in fact there are only finitely many twin primes. That proof is factive just means that for this to be true – for them to have really proved that there are only finitely many twin primes – it must in fact *be* the case that there are only finitely many twin primes. It is nonsense to suggest that their proof might

[1] The most significant difference is that ETCS lacks replacement and has separation for only bounded formulae, but one can address these issues by adding a replacement axiom scheme, as McLarty (2004) discusses.

be valid, but that the number of twin primes might still be infinite. This view of proof can be evidenced by remarks by mathematicians, and commentators on mathematics. For instance:

> for the mathematician, truth is established via proofs (Aschbacher 2005, p. 2403)

> Mathematicians know that a formal proof leads always to a correct result (Atiyah et al. 1994, p. 200)

> *Proof* is our device for establishing the absolute and irrevocable truth of statements of mathematics (Krantz 2011, p. 3).

Though this is I think the way we naturally think of proof, it presents a problem for mathematics. If a mathematician proved that there were between 10^{100} and 10^{200} twin primes, then their conclusion would seem to imply that prime numbers exist, and in particular that natural numbers exist. Though in my experience mathematicians are often reasonably happy to assert the existence of the natural numbers, they may well be less sure of this than they are of the correctness of any particular proof about the natural numbers (especially a straightforward one). And many mathematical proofs concern structures much more esoteric than the natural numbers, whose existence mathematicians will be more reluctant to categorically assert.

Thus though many mathematicians do believe in the reality of the objects and structures they study, mathematicians often officially retreat to a reinterpretation of mathematical assertions that does not rely on the existence of mathematical entities. One popular paraphrase – the if-thenist option – takes a mathematician's assertion that ϕ to mean that if the relevant axioms are true, then ϕ. For instance when a mathematician says that there are finitely many twin primes, we might interpret them as meaning that if the axioms of arithmetic are true, then there are finitely many twin primes. A similar

paraphrase, which we could call consequentialist, takes the assertion that ϕ to mean that ϕ is a logical consequence of the relevant axioms (both these options are described in more detail in section VI.3). This is described by Gowers (2002, p. 41) as often appealed to by mathematicians. Other less definite attitudes, such as a feeling about how a certain kind of object behaves, or a hunch about the likelihood of a particular conjecture, could be paraphrased in similar ways. The retreat to this kind of paraphrase may be alluded to in the aphorism that mathematicians are Platonists on weekdays, and formalists on Sundays (Hersh 1997, pp. 39–40; Corry 2012, p. 310): really believing in the reality of their subject matter while investigating it, but appealing to a perhaps formalist paraphrase when this reality is questioned. It is important to realise that the use of these kinds of reinterpretations is not in conflict with the genuine, factive conception of proof: to the contrary, it is because mathematicians wish to be assured that the conclusions of their proofs are true that they retreat to restating the content of these conclusions in these if-thenist or consequentialist terms. As seen in section I.7 and mentioned above, one can argue that an informal proof in T is formalizable in principle according to the rules of T, and thus (if the logic of T is classical) establishes that its conclusion is a logical consequence of the axioms used: thus if one proves results informally in T, both the if-thenist and consequentialist interpretations relative to the axioms of T are straightforwardly factive.

In general the factive property of proof (in the genuine or epistemic sense) requires that an argument p can only be a proof of ϕ if ϕ is in fact true. To genuinely prove a statement ϕ, we thus need some sort of interpretation of what it is for ϕ to be true or false. The if-thenist and consequentialist views just mentioned can be viewed as interpretations of this kind, whose primary purpose is to make proof straightforwardly factive.

3 Interpretations of mathematics

We now discuss a variety of interpretations of mathematical statements that have been put forward, and how they relate to the issue of factivity of proof raised in the previous section. After the initial clarification of if-thenism, consequentialism and the related view of deductivism, much of the remainder of the section consists of exposition of interpretations that one who knows the field will already be familiar with. How the question of factivity plays out for these different interpretations is also considered though, and a summary of this is located at the end of the section.

Firstly, we can say a little more about the if-thenist and consequentialist interpretations. Each takes there to be an ambient set of axioms, with the if-thenist interpreting ϕ as meaning something like "if the axioms are true, then ϕ", and the consequentialist interpreting ϕ as meaning something like "ϕ is a logical consequence of the axioms". In both cases, we will talk about the interpretation being relative to an axiom set – and which axiom set this is turns out to be very important.

Though one can roughly characterize these interpretations in the above manner, if one tries to make them more precise then there are various different options. Suppose for instance that we have a statement ϕ of arithmetic. An if-thenist might interpret ϕ as really meaning that there exists a finite subset Γ of the axioms of PA such that $(\bigwedge_{\gamma\in\Gamma}\gamma)\rightarrow\phi$ (this being a material implication, or perhaps a natural language "if...then"). Otherwise we could give a second order interpretation in terms of second order arithmetic PA^2, interpreting ϕ as meaning that $(\bigwedge_{\gamma\in\mathrm{PA}^2}\gamma)\rightarrow\phi$. These will be referred to as "local" if-thenist interpretations. They are to be contrasted with "global" if-thenist interpretations, where we take a reduction of ϕ to a statement ϕ^{T} of some general axiom system T such as ZFC (for instance translating ϕ into a statement about a particular set theoretic simply infinite sequence), and then interpret ϕ as the statement that there are finitely many axioms γ of T such that $(\bigwedge_{\gamma\in\Gamma}\gamma)\rightarrow\phi$. Alternatively, as in the local case,

one could use a second order formulation of the axiom system (such as Morse-Kelley set theory). Similarly for instance if ϕ states a property of rings, one can interpret it either locally as the statement that the conjunction of the ring axioms implies ϕ, or globally by reducing ϕ to a statement ϕ^{T} of some general axiom system, and interpret ϕ as meaning that some finite conjunction of axioms of T implies ϕ^{T}. How to give a local if-thenist interpretation of a statement about for instance manifolds or schemes or some other mathematical kind that we do not have a natural intrinsic axiomatization for is less clear – though we can often find reasonably self contained axiomatizations, as discussed in section VI.4. There are similar distinctions to be made in the consequentialist case. If ϕ is a statement of arithmetic, it can be interpreted locally as $\mathrm{PA} \vdash \phi$, or $\mathrm{PA}^2 \vdash \phi$, or globally as for instance $\mathrm{ZFC} \vdash \phi^{\mathrm{ZFC}}$ where ϕ^{ZFC} is a reduction of ϕ to set theory.

We can denote different interpretations of mathematics with superscripts.[2] For instance we could denote the local if-thenist interpretation of statements ϕ of arithmetic by $\phi \mapsto \phi^{\mathrm{If}_{\mathrm{PA}}}$, and the global interpretation with respect to ZFC by $\phi \mapsto \phi^{\mathrm{If}_{\mathrm{ZFC}}}$. The corresponding consequentialist interpretations could be denoted by $\phi \mapsto \phi^{\mathrm{Cons}_{\mathrm{PA}}}$ and $\phi \mapsto \phi^{\mathrm{Cons}_{\mathrm{ZFC}}}$.

Awodey (2004) appears to defend something like an if-thenist interpretation of mathematics, though he does not distinguish between the different axioms sets one might have in mind, or make it clear whether he intends his if-thenism to be local or global.[3] These distinctions are important, as we will see.

A first sign of their importance concerns the factivity of proof. As discussed in section VI.2, if T is a proof system then the if-thenist and consequentialist interpretations relative to T are straightforwardly factive with regards to proof in T. For instance, a purely arithmetic proof of ϕ immediately establishes that there exists a finite subset Γ of the axioms of PA such that $(\bigwedge_{\gamma \in \Gamma} \gamma) \to \phi$, and that $\mathrm{PA} \vdash \phi$. The notion of proof

[2] I am not claiming the notation suggested here is the most attractive possible.

[3] He generally talks in terms of schematic statements about kinds of structures, but mentions at one point (Awodey 2004, p. 60) that the axioms of a system such as ZFC may be conventionally assumed.

in a general axiom system such as set theory is not necessarily factive with respect to local consequentialist interpretations, however: if we have a proof of ϕ in set theory, it does not follow that PA $\vdash \phi$ (for instance taking ϕ to state the consistency of PA). The case of the local if-thenist interpretation is less clear, and might depend on how one interprets the relevant vocabulary and the conditional. For instance one might be able to argue that if ϕ is a statement of arithmetic provable in set theory, then it is true; and thus that taking Γ to be just the empty set, we have $(\bigwedge_{\gamma\in\Gamma}\gamma) \rightarrow \phi$. Proof in set theory would then be factive with respect to the local if-thenist interpretation.

In some ways it is a pity that mathematicians so often opt for the if-thenist or consequentialist interpretations, since there are more attractive versions of both. Similar to consequentialism is what might be called the deductivist interpretation of mathematics. This approach is the one taken by Tait (2005) and Muller (2004), mentioned in the introduction. They argue from the Wittgensteinian conception of meaning as use that any consistent axiom system gives a valid sense to its vocabulary, with the rules of the axiom system governing when its statements can be asserted or denied. The difference between this and consequentialism is when we assert an existential mathematical statement, for instance, consequentialism takes us to be talking metalinguistically, talking about our statement and asserting that it follows from certain principles: deductivism, by contrast, takes the existential statement to be literally valid, but with its meaning given by the deductive rules for the existential quantifier (such as quantifier introduction and elimination) in the system. This difference is subtle, but is the difference between a formalist interpretation of mathematics and one on which it has genuine content.

Of course mathematicians generally do not work explicitly with the formal rules of an axiom system, but as discussed in chapter I their arguments can be seen as more compressed, higher level versions of arguments that do. Statements in informal mathematical proofs can thus be seen as valid assertions, gaining a sense either from the norms governing high level proof or from the underlying axioms – in the same way as one

can see the meaning of natural language statements involving conjunction or disjunction as derived from the formal rules for these connectives, though these formal rules might not actually be explicitly used in speech. Thus we obtain an interpretation which takes the sense of mathematical statements to be given by the norms for proving or refuting them. This interpretation trivially makes proof factive (for proofs in the relevant axiom system, at least).

Deductivism comes in various different forms. On the view of Tait (2005), we are apparently free to extend our axiom system in an open ended way, adding new axioms as we see fit, as long as we retain consistency.[4] We could denote this by $\phi \mapsto \phi^{\mathrm{D_{oe}}}$. Muller (2004) by contrast seems to have a view where we determine the meaning of a mathematical concept (such as "set") by deciding on a fixed list of axioms. This might give us an interpretation $\phi \mapsto \phi^{\mathrm{D_{ZFC}}}$ of statements reducible to set theory and $\phi \mapsto \phi^{\mathrm{D_{PA}}}$ of statements reducible to first order arithmetic. Finally Maddy (2011) can also be regarded as giving a kind of deductivist interpretation of mathematics, with the norm for asserting mathematical statements given by deducibility in set theory, but in this case the set theoretic axioms are regarded as open ended in a more constrained way than by Tait – with us choosing whichever new axioms best increase mathematical depth and fruitfulness (properties she regards as reasonably objective). We might denote this by $\phi \mapsto \phi^{\mathrm{D_{md}}}$.

Just as deductivism is perhaps a more attractive cousin of consequentialism, the view known as eliminative structuralism is a more flexible version of local if-thenism. On this interpretation, if ϕ is a statement about the natural numbers, instead of interpreting it to mean "if the axioms of arithmetic are true, then ϕ", we interpret it to mean roughly "for any structure, if the axioms of arithmetic hold of that structure, then ϕ does". More formally, we axiomatize (in second order logic, say) the notion of a relation R being the

[4]For the open-endedness of the axioms see Tait (2005, pp. 91, 96–98, 294–295). That there are no objective norms for correctness when extending an axiom system (beyond human considerations of naturalness or fruitfulness for instance) comes out at Tait (ibid., p. 97) and the contrast with Gödel's views at Tait (ibid., p. 294).

successor relation of a simply infinite sequence, as Inf-Seq(R).[5] Then, writing $\phi[R]$ for the reconstrual of a statement of arithmetic as about the infinite sequence defined by R, we interpret such a ϕ as meaning

$$\forall R \ (\text{Inf-Seq}(R) \rightarrow \phi[R]).$$

This approach uses the fact that we can often find "real world" characterizations of concepts (such as infinite sequence) usually defined formally in a mathematical axiom system. "Real world" here means that such a characterization can be satisfied, or fail to be satisfied, by real world objects – things that "exist" according to the normal meaning of that term (though what exactly does exist is of course a contentious issue). This ability to find real world characterizations of formal concepts is discussed in more detail in section VI.4, and applies to many other kinds of structures than just infinite sequences – for instance we can interpet statements about the real numbers as being about all complete ordered fields, as axiomatized in second order logic, and interpret statements about all groups in terms of all ternary relations R that define a group structure. We can denote this interpretation by $\phi \mapsto \phi^{\text{Elim}}$. For this interpretation, the question of whether proof is factive is finally an interesting one, as it is for the following structuralist interpretations. Indeed, how would one argue for a given theory T that proving statements of arithmetic in T actually delivers truths about simply infinite sequences that exist? What about proofs about other kinds of structures? This is essentially the question of soundness for T, as defined in section VI.6.

The standard objection to this interpretation is that it may prove to be vacuous – if no simply infinite sequences exist, for instance, then all statements of arithmetic

[5] By "simply infinite sequence" I mean the standard notion with initial element and successor operation – to be distinguished for instance from the notion of a sequence as a function $\mathbb{N} \rightarrow X$ for some X, or a sequence with length some larger ordinal. The axioms for this notion state that R defines an injective function on its domain, and that its domain contains a unique element not in its image, and that its domain is the smallest collection of objects containing this unique element and closed under the operation of applying R.

come out as trivially true. If one accepts the language of necessity and possibility as meaningful, then one can move from eliminative structuralism to modal structuralism, defended by Hellman (1993). On this interpretation, mathematical statements are not just about all structures that do exist, but all structures that could exist. Using modal logic, a statement ϕ of arithmetic is interpreted as saying for instance that

$$\Box\ \forall R\ (\text{Inf-Seq}(R) \rightarrow \phi[R]).$$

There is much less threat of vacuity on this interpretation than the previous one, since the claim that a simply infinite sequence could exist is much weaker than the claim that they do exist. Exactly how much mathematics comes out as non vacuous will depend on the modality used: the notion of logical possibility covers more structures than that of physical possibility, potentially allowing a non vacuous interpretation even of set theory. We can denote this interpretation by $\phi \mapsto \phi^{\text{Modal}}$.

The final main structuralist interpretation of mathematics is *ante rem* structuralism, defended by Shapiro (1997). This view holds that the natural numbers structure exists as the form of simply infinite sequences, a kind of abstract pattern, even if no other simply infinite sequences do. The individual natural numbers $0, 1, 2, \ldots$ are the "places" of this structure, universals in a sense, corresponding to the various positions (0^{th}, 1^{st}, 2^{nd}, ...) of other infinite sequences. Arithmetic is then interpreted as about this particular distinguished infinite sequence. Statements about the real numbers are interpreted in the same way as about the structure which is the form of complete ordered fields, and similarly for statements of other categorical theories. For non categorical theories such as group theory, there are various different group structures – the forms of each isomorphism type of groups – and generalizations about groups are interpreted as about all these group structures. We can denote this interpretation by $\phi \mapsto \phi^{\text{ante}}$.

Apart from structuralist views, the most discussed philosophical interpretation of

mathematics has been Neo-Fregeanism (Hale and Wright 2004; Wright 1983) This view sees the meaning of mathematical vocabulary as given by abstraction principles, for instance

> the number of F's is equal to the number of G's iff there is a bijection between the F's and the G's

as a principle governing how identity for number terms ("the number of F's" and the like) behaves. Neo-Fregeans have sought similar abstraction principles for other mathematical entities, with some success for the real numbers (Hale 2005) but less for the further reaches of set theory. This view could be considered somewhat similar to the deductivist view, where the meaning of mathematical vocabulary is given by the rules for reasoning about it; the main difference is that the numerical vocabulary Neo-Fregeans see as introduced in this way is directly applicable to the real world, and may allow us to speak and reason more powerfully about it, whereas on the standard deductivist view the mathematical vocabulary is semantically isolated from other areas of discourse. We can denote the Neo-Fregean interpretation by $\phi \mapsto \phi^{\text{Neo-F}}$. Again, the question of whether proof in a general system T is factive is a substantial one on this view.

Another interpretation that should be mentioned is fictionalism. First proposed by Field (1980; 1989), this view holds that the face value Platonist interpretation of mathematics is correct, but that all statements thus interpreted are false since the abstract objects being quantified over do not in fact exist.[6] We can denote this by $\phi \mapsto \phi^{\text{fiction}}$. On the fictionalist interpretation, essentially by definition, proof is never factive. Thus though a fictionalist can recognize the notion of a (formal or informal) derivation in T, for a proof system T, they can never regard mathematicians as genuinely proving anything. One can try to take Field's approach, which takes an instrumentalist attitude to mathematics, regarding it as a useful tool for deducing extra physical facts, but still

[6]There has been substantial subsequent arguments by fictionalists, such as by Balaguer (2001) and Leng (2010), though these authors have generally focused on defending fictionalism from the indispensability argument rather than giving positive arguments in favour of fictionalism.

cannot regard mathematical arguments as genuine proofs. This I think makes fictionalism an unattractive option, when lightweight realist views such as Tait and Muller's deductivism are available – so that for fictionalism to be viable, it needs to be buttressed by an argument as to why the substantial (but false) interpretation of Platonism it advocates is the correct one, instead of these more minimalist views.

A final possible interpretation of mathematics worth mentioning is set theoretic realism. This comes in many forms, whose common ground resides in regarding set theoretic statements as a real description of some portion of reality. One might believe that the set theoretic hierarchy exists, formed by successive applications of the set formation operator (for instance Boolos 1971 and Burgess 2004); or one might take a potentialist view based around the idea that any collection of objects could have formed a set (for instance Linnebo 2010; 2013 and Studd 2013). On these views we can only correctly postulate set theoretic principles which actually hold of the hierarchy – we are not free to stipulate the properties of sets, as one may be on a deductivist view like that of Tait (2005) or Muller (2004) discussed above. Given the sets, one can then interpret mathematical statements in set theory in the usual way: for instance interpreting statements of arithmetic in terms of some particular set theoretic infinite sequence, such as the finite Von-Neumann ordinals or the sequence $\{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}, \ldots\}$, interpreting statements of group theory as about all set theoretic group structures, and so on. One could denote such an interpretation by $\phi \mapsto \phi^{\mathrm{Set}}$. On these interpretations whether proof in set theory is factive will come down to whether the set theoretic principles we postulate do actually hold of the hierarchy of sets.

There are various other notable interpretations of mathematics that have been put forward, such as the interpretation of Chihara (1991) in terms of constructibility, of Kitcher (1985) in terms of ideal agents, and Balaguer (2001) in terms of plenitudinous Platonism. It isn't possible here to consider all interpretations that have been defended however, and how the questions of factivity and soundness apply to other interpretations

is often similar to how they apply to certain of the interpretations above.

Finally, to summarize the discussion of factivity of proof with regards to the views considered. The if-thenist and consequentialist views are defined relative to some axiom system, and proof in this axiom system is then automatically factive; but proof in other proof systems need not be factive, with for instance proof in ZFC not necessarily being factive for a consequentialist view with respect to the axioms of PA (which interprets ϕ as meaning that ϕ is a logical consequence of the axioms of PA). Deductivist interpretations come in different forms, but all are relative to a choice of proof system, with proof in this proof system again then automatically being factive – but proof in other proof systems perhaps not being factive. For the various structuralist views, and for Neo-Fregeanism, the factivity of proof in a proof system like ZFC can be a difficult question: for instance the question of whether if ZFC proves a statement of arithmetic, that statement of arithmetic actually holds of all simply infinite sequences, or of the distinguished simply infinite sequences picked out by *ante rem* structuralism or Neo-Fregeanism. Then there is fictionalism, for which – essentially by definition – proof is never factive, so that there can be no notion of genuine proof in mathematics. As discussed above, this seems to be an unattractive feature of fictionalism. Finally, interpretations of mathematics in set theory, which take seriously the existence of the set theoretic hierarchy, were discussed – with the key issue for factivity here being whether the axioms of set theory we postulate are actually true of the set theoretic hierarchy.

4 Realizations of mathematical concepts

Mathematical concepts are typically officially defined within the context of a mathematical proof system; for instance, the concept of group can be defined in ZFC as an ordered pair (G, m) where G is a set and $m : G \times G \to G$ is a function which satisfies the group axioms (associativity, existence of an identity, and existence of inverses). However this

section lays out how even when a concept is officially defined in such terms, we can still often find a real world characterization of "the same" concept – where by a real world characterization we mean one which can be satisfied, or fail to be satisfied, by real world objects. Such a characterization can be independent of any broader mathematical proof system, and in many cases can be more or less self contained, not relying on any objects outside of the structure of interest. This can be seen as a counter to the view of Maddy (2011, p. 92), in which mathematical objects and structures always come saddled with whatever properties the background set theory gives them. The work done in this section will be used as the basis for a condition that section VI.5 argues we should require of acceptable interpretations of mathematics.

As noted initially, mathematical concepts – like that of a group – are generally officially defined within the context of a mathematical proof system, such as set theory. Maddy takes these kinds of set theoretic definitions very seriously, stating that

> [a] mathematical structure ... doesn't exist in a vacuum; it's embedded in a rich mathematical universe – V if you like – and it has all the properties the [methods of set theory] reveal ... as part of its identity as a mathematical object (ibid., p. 92)

(Maddy is speaking of a particular example, but her comment is intended as a general one). On her view, the status of mathematical structures as elements of a broader set theoretic universe – with the properties set theory discerns in them – is essential to them.

We can start developing a different perspective by noting that the exact form that such a set theoretic definition takes is generally mathematically irrelevant. For instance, one could otherwise just define a group to be a function $m : G \times G \to G$ for some set G which has the appropriate properties, or a quadruple (G, m, e, i) where G and m are as above and e is the identity for m, $i : G \to G$ the map giving inverses. All these capture "the same" mathematical concept, despite the formal differences: mathematically, we can do the same things with all of them. We can also capture "the same" concept in

separate mathematical proof systems: for instance defining a group in the categorical theory ETCS as a triple of arrows $m : G \times G \to G$, $e : 1 \to G$, $i : G \to G$ such that the appropriate diagrams commute; or using the correspondence between propositions and types in homotopy type theory to give a definition of group as a particular type (Univalent Foundations Program 2013, §1.11).

All of these definitions are different ways of coding the same information. To reason about groups we need to know what we can do with their elements – multiply them, or take inverses – and what properties these satisfy. We can think of these various definitions of group as being different ways to formally capture the same informal mathematical concept.

The same applies much more widely: when we make a formal definition of a mathematical concept, we can think of there being an informal concept behind it – which a variety of potential formalizations could adequately capture.[7] The point of this section is to argue (*pace* Maddy) that such informal concepts can often also be captured by a characterization independent of any substantial mathematical proof system, and further, by a characterization that has real world content – allowing the possibility of real world instantiations of the formal concept, which we will also refer to *realizations* or *real world examples* of it.

Suppose for instance that you are sitting on a chair c with a laptop l in front of you. One can define a ternary relation R, where we have $R(c, c, c)$, $R(c, l, l)$, $R(l, c, l)$ and $R(l, l, c)$, and no other values related by R. It is immediate that this acts as a binary operation in its first two arguments, with domain consisting of c and l – so that as long as x and y are amongst c and l, there is a unique z such that $R(x, y, z)$. It is straightforward to check that this binary operation satisfies the group axioms – it is associative, c is an identity element, and c and l are each inverses of themselves. Thus, morally, R makes your chair and laptop into a group.

[7] This parallels the view of Awodey (2004), who regards mathematical generalizations as really being schematic assertions that can apply to a variety of different formal concepts.

This is despite the situation being little like that described by the ZFC definition of group. Under the (G, m) definition a group is literally an ordered pair, which is literally – on the usual definition – the set $\{\{G\}, \{G, m\}\}$, with m being in turn a set of ordered pairs. Thus for l and c to form a group one would have to believe in the real ability to form not just sets of everyday objects, but sets of sets (and sets of sets of sets, and so on) of such objects. I do not mean to claim that this is implausible, just that it is a further assumption, beyond anything involved in the above case of R. Moreover, in standard ZFC there is no room for distinct objects such as c and l which aren't sets – since by extensionality, if two objects both have no members then they are equal.

None of this prevents us from regarding the relation R as defining a kind of group structure on the chair and laptop. There is nothing mysterious about the nature of this R – this is just a variable standing for an open formula of our language, which we could write as

$$(x = c \wedge y = c \wedge z = c) \vee (x = c \wedge y = l \wedge z = l) \vee$$
$$(x = l \wedge y = c \wedge z = l) \vee (x = l \wedge y = l \wedge z = l)$$

with arguments x, y and z. We can then reason about the properties of such relations in second order logic; as discussed in Appendix B, as long as we limit ourselves to restrictive predicative second order logic (in which the comprehension scheme is restricted to open formulae lacking second order variables) this does not commit us to any abstract entities, or indeed any entities beyond the open formulae we actually write down. We can define a notion of group in this logic, stating as a formula $\text{Group}(P, R_{\times})$ what it is for predicate P and ternary relation R to "be" a group, with $\times$ defining a binary operation on the objects falling under P satisfying the properties discussed above. We can take an open ended view of the quantifiers of this logic, so that if we assert a universal statement about these group structures we are committed to that statement still holding no matter how

the vocabulary of our language expands in future, potentially allowing us to define new groups. Of course one can axiomatize the notion of group in first order logic, but using this fragment of second order logic allows us to state and argue for generalizations about all groups, so defined.

Though – as just discussed – this notion of group is formally quite distinct from the definition in ZFC, it is faithful to the informal concept underlying the definition: we know what we can do with elements of groups (multiply them and take inverses), and what properties these have. It is also what we can call a real world characterization: it is the kind of characterization that can be satisfied, or fail to be satisfied, by real world objects (and relations between them). The objects being quantified over when we use this kind of logic are just the objects that genuinely "exist", whatever we mean by that word; the logic is just a regimented, clarified version of a fragment of natural language, with the range of the first order variables just being the range of the quantifiers "there are" and "for all" in English, including things such as laptops and chairs (the question of what objects are properly included under these quantifiers is of course a vexed one). Thus this notion of group is what we can call a faithful real world characterization of the formal concept of group. We will call a relation satisfying it a *realization* of the formal concept of group: a real world instantiation of this formal concept. This is a realization of the formal concept of group however the formal concept is defined – in ZFC, or ETCS and so on – with all these formal concepts having the same underlying informal concept behind them.

The same goes for other concepts that can be axiomatized in first order logic. A graph would commonly be defined as a pair (V, E) where V is a set and every $e \in E$ is of the form $\{u, v\}$ with $u, v \in V$, $u \neq v$. Then a faithful real world characterization of this is just given by a predicate (one place relation) V and a binary relation R which is antireflexive and symmetric and only relates things falling under G – so that G determines the vertices of the graph, and R determines its edges. We call any such V and

G a realization of the formal concept of graph. For one such realization we could take the V to apply to the employees of a company, and take the relation $R(x, y)$ defined by x and y being employees that are facebook friends (assuming there are no complications like multiple accounts).

There may be various plausible options for a faithful real world characterization of a formal concept. For instance one might take a group to be given by a predicate G as well as a ternary relation R with the properties discussed above, where G applies to the objects making up the group. One may argue about which such variations are faithful, or the most faithful, but this has little importance to any of the following.

One can obtain similar characterizations of other kinds of structures axiomatizable in first order logic, such as monoids, rings, fields, partial orders, and categories (the use of restricted predicative second order logic here is to allow us, in effect, to state generalizations about such structures). We can also characterize structures axiomatizable in second order terms. There are different logical options for doing this. If we supplement restricted predicative second order logic with plural logic (the interaction of other logics with restricted predicative second order logic is discussed towards the end of Appendix B), then we can give a familiar characterization of a binary relation R being the successor function of an infinite sequence made up by the objects satisfying predicate P – where R has these objects as its domain, and this domain contains a unique "initial element" not in the image of R, and this domain is the smallest plurality containing the initial element and closed under the operation of applying R (the smallest plurality being the intersection of such pluralities). We will call the last clause here the inductive clause. This is a faithful characterization of the notion of simply infinite sequence, paralleling the usual set theoretic definition, except with quantification over pluralities in the inductive clause instead of quantification over subsets.

In fact there is a case to be made that restricted predicative second order logic alone is enough for a faithful characterization, provided we have an open ended conception

of the range of predicate variables – expanding the range of the comprehension scheme in restricted predicative second order logic to include new formulae definable as the vocabulary of our language expands. Indeed we can give a similar characterization in restricted predicative second order logic, except with the plural logic inductive clause replaced by the condition that the entire domain of R falls under any predicate containing its initial element and closed under the operation of applying R. Then one can argue that this is in fact equivalent to the characterization using plural logic. All we have to show is that the two inductive clauses are equivalent. That an R satisfying the plural logic inductive clause also satisfies the restricted predicative second order logic inductive clause is straightforward, using comprehension for pluralities (and the ability to use free predicate variables in the plural comprehension scheme, as discused in Appendix B). For the converse, given an R we can let xx be the smallest plurality containing its initial object and closed under the operation of applying R, and can then argue that this plurality consists exactly of those objects in the domain of R, by putting the predicate "is among the xx's" into the inductive clause of the restricted predicative characterization (as discussed in Appendix B we allow free plural variables in instances of restricted predicative comprehension), and noting that the plurality of objects in the domain of R is a plurality containing R's initial object and closed under the operation of applying R. The case of full second order logic is similar, if one allows two types of second order variables, the first with impredicative comprehension and the second with the restricted predicative comprehension scheme (in which second order variables of the first type are allowed, as discussed in Appendix B).

This characterization of simply infinite sequence just involves the notion of successor function – essentially the operation of "adding 1" – rather than the more complex arithmetic operations of addition and multiplication. These can be defined in terms of the successor function as long as one has the resources to define primitive recursive functions. One can do this in full second order logic, with its more powerful comprehension

scheme for relations (actually predicative second order logic is sufficient); however our preferred option is to use double ancestral logic, which was defended as a valid logic in chapter IV, and which naturally captures the ability to define primitive recursive functions. Either way, if we have the resources to define primitive recursive functions, then we can obtain addition and multiplication operations on any simply infinite sequences.

By combining both double ancestral logic and plural logic with restricted predicative second order logic, we can prove a categoricity result for simply infinite sequences. Indeed given two infinite sequences with successor relations R and R', and initial objects a and a', we can define a primitive recursive isomorphism f from the first to the second – which maps a to a', and where f applied to the R-successor of x is the R'-successor of f applied to x.[8] This is an instance of the ability to define general primitive recursive functions seen in section IV.4, though in a setting with different logics in play. Plural logic is used to establishing that this is an isomorphism (which could otherwise be done by using the ancestral, definable in terms of the double ancestral): for instance we can form the plurality of objects which are mapped by f to something lying in the domain of R', and argue by induction along R that this contains every element in the domain of R, and thus that f is total. Here we use the ability to use the free variable R' in instances of plural comprehension, which is not possible merely with restricted predicative comprehension.

The case of complete ordered fields is similar. We can give a faithful characterization by stating what it is for relations R_+, $R_\times$ and $R_<$ to give an ordered field structure on the objects falling under a predicate P, and stating completeness as the property that for nonempty pluralities of objects falling under P, if there is an $R_<$ upper bound for those objects then there is a least upper bound. Again by combining both double ancestral logic and plural logic with restricted predicative second order logic, we can

[8]The result here is of the form "for all simply infinite sequences R and R' the open formula $\chi(x, y)$ defines an isomorphism from R to R' such that..." where χ contains R and R' free. We cannot move from this to the $\forall\exists$ statement "for all simply infinite sequences R and R' there is an isomorphism from R to R' such that...", since as χ contains R and R' free it is not suitable for use in an instance of restricted predicative comprehension.

prove a categoricity result for this characterization of complete ordered fields – given two such complete ordered fields, obtaining an isomorphism first between their natural numbers, extending this to their integers, then their rationals, then all their elements using the completeness property.

The possibility of characterizations of this kind is a counter to Maddy's view of mathematical structures, in which their status as elements of the universe of sets is essential to them. Of course, as discussed initially, mathematical structures typically are (officially) defined in set theoretic terms; but this need not prevent us from finding alternative characterizations of the "same" mathematical concept independent of set theory. This is particularly straightforward for structures axiomatizable in first or second order terms, as discussed. In these cases, the only objects involved in the characterizations are elements of the structure we are interested in characterizing. This need not always be the case. For instance one could characterize the notion of real vector space by using the above characterization of predicate P and relations R_+, $R_\times$, $R_<$ giving a complete ordered field structure, and stating what it is for S_+ and $S_\times$ to define the addition and scalar multiplication operations of a vector space structure on objects falling under predicate Q with scalars in this complete ordered field. In this case our real structure of interest is the vector space itself, but to characterize it we are led to use the elements of a complete ordered field as auxiliaries. We will talk about such a characterization *involving auxiliaries*, with characterizations that don't involve auxiliaries (like the other ones considered so far) being *self contained*. On Maddy's conception of mathematical structures, their place in the set theoretic hierarchy is always essential to them; if whenever we tried to characterize a structure we needed the whole set theoretic hierarchy as auxiliaries, this would be a plausible view, but using the whole set theoretic hierarchy as auxiliaries is, it appears, very rarely necessary.

We have already seen some examples of real world instantiations of simple mathematical structures, such as finite groups and graphs, but there may be a worry about

whether more complex, substantial mathematical structures will also be instantiated. Maddy expresses a concern somewhat along these lines, arguing that when applying mathematics,

> we aren't in fact uncovering the underlying mathematical structures realized in the world; rather, we're constructing abstract mathematical models and trying our best to make true assertions about the ways in which they do and don't correspond to physical facts. There are rare cases where this correspondence is something like an isomorphism ... but most of the time the correspondence is something more complex, and all too often it's something we simply don't yet understand (Maddy 2011, pp. 27–28)

This I think is a perceptive and accurate description of the way applied mathematics currently works in practice, which is what Maddy – as a committed naturalist – is aiming for. However these kinds of observations should not be taken as any kind of guarantee that substantial mathematical structure does not, in fact, exist in the world (Maddy does not claim this, though she does conduct her discussion as though the possibility is irrelevant).

For instance it may be true that current physics does not confirm the existence of infinite sequences in the world, or place importance on them, but that does not mean that it rules them out. It appears for instance that the question of how the universe will end is still undecided by physics. One possibility is that it may last forever, ending in a drawn out heat death; otherwise it may collapse in finite time in a Big Crunch; otherwise it may perhaps continue forever in an infinite sequence of Big Crunches followed by Big Bangs. If it does last forever, then we can straightforwardly obtain a real world instantiation of the notion of simply infinite sequence – for instance taking the sequence of Big Crunches, if there are infinitely many of them, or just a sequence of spacetime points evenly spaced along some trajectory if not. If the universe will only last for finite time, but it is spatially infinite, one could similarly obtain a simply infinite sequence

by picking out a non-intersecting path and taking points evenly spaced along it. Even if the universe is spatially and temporally finite, space and/or may be continuous – Maddy (1997, pp.146–152) considers evidence that they are discrete, but this is far from definitive – in which case we can select an arbitrary point and obtain a simply infinite sequence of points approaching it. In all these cases we are not finding some sort of complex correspondence between part of the world and a set theoretic simply infinite sequence – the kind of outcome Maddy discerns as typical in applications of mathematics – nor even merely an isomorphism: we are literally finding the structure of a simply infinite sequence instantiated in the world, as characterized above in a logic suited to the purpose.

The same circumstances that allow the definition of a real world infinite sequence generally also allow the definition of two disjoint sequences, in which case we can combine the two – one to play the role of the non-negative integers, the other the negative integers – and using the ability to define primitive recursive functions in double ancestral logic, can obtain real world groups and rings isomorphic to $\mathbb{Z}$. This gives more substantial groups and rings than simple finite ones like that considered initially.

The structure of a complete ordered field may seem like one that the physical realm is unlikely to instantiate, but with a slight shift in perspective we can see how it may arise naturally. It is a basic theorem of order theory that the real line is (up to isomorphism) the unique nonempty dense separable complete total order without endpoints. This list of conditions is far from arbitrary: it is a good candidate for what we mean by continuity of an open interval in time, or continuity of an ordered arc we pick out in space or spacetime (perhaps just ordered along the rough direction of the arc). Using restricted predicative second order logic and plural logic together we can state what it is for relations R_S and $R_<$ to give such a structure, which we term a continuously ordered open interval: we require that $R_<$ is a complete dense total order without endpoints on some nonempty domain, and that R_S is the successor function of a simply infinite sequence which takes

values in this domain and which is dense under the ordering $R_<$. We state both the completeness property of $R_<$ and the inductive clause of R_S in plural logic. Appendix C then proves in this setting that a continuously ordered open interval can have a field structure defined on it, with respect to which it becomes a complete ordered field (this uses double ancestral logic in addition to plural logic and restricted predicative second order logic).[9] Since – as noted above – the continuity of space and time is an open question, there is every possibility that there might be physical continuously ordered open intervals (if realism about moments in time or space/spacetime points is correct, at least), in which case we would thus obtain physical realizations of the structure of a complete ordered field.

Though simply infinite sequences and complete ordered fields are perhaps the most totemic, the variety of structures that mathematics investigates is of course immense. It will be instructive to consider what real world characterizations might look like in a few other cases, though there is no possibility of answering this question for all the different kinds of mathematical structures.

Firstly, we have the concept of a topological space, usually defined as a set X equipped with a topology – a set of "open subsets" of X, which contains X and the empty set, and is closed under unions and finite intersections. Though this is phrased in explicitly set theoretic terminology, the key feature is just that we have an appropriate membership relation between elements of X and the "open subsets", with the ability to take unions and intersections of the latter. This situation can be straightforwardly axiomatized using restricted predicative second order logic together with plural logic.

[9] As with the case of categoricity for simply infinite sequences discussed above, all we obtain here are open formulae $\chi_+(x, y, z)$ and $\chi_\times(x, y, z)$ which contain $R_<$ and R_S as free variables, and which define addition and multiplication operations which together with $R_<$ give a complete ordered field structure. Due to the limitations of the logic, we are unable to move from this to a $\forall\exists$ statement that for any $R_<$ and R_S there are R_+ and $R_\times$ which which define an appropriate field structure, since this would require the ability to state instances of comprehension with χ_+ and $\chi_\times$, which is ruled out in restricted predicative second order logic due to the presence of free second order variables. However if we work with open formulae $\chi_<(x, y)$ and $\chi_S(x, y)$ in place of $R_<$ and R_S then the resulting χ_+ and $\chi_\times$ do not contain free second order variables, so we can deduce the appropriate existential conclusion via comprehension.

We need two predicate variables P and P_{open}, and a relation variable $R_{\in}$ where $R_{\in}(x, y)$ holds only if $P(x)$ and $P_{\text{open}}(y)$, where we think of $R_{\in}(x, y)$ as stating that x is a member of y. Then we state extensionality for our "open sets", that if $P_{\text{open}}(y)$ and $P_{\text{open}}(y')$ then $y = y'$ iff for all x, $R_{\in}(x, y)$ iff $R_{\in}(x, y')$. Finally we state that the analogues of $\varnothing$ and $\{x \mid P(x)\}$ are open, and that the open sets are closed under binary intersections and arbitrary unions, in the obvious way – for instance stating that there is y such that $P_{\text{open}}(y)$ and $R_{\in}(x, y)$ iff $P(x)$, and that if yy is any plurality of objects all falling under P_{open} then there is z such that $P_{\text{open}}(z)$ and $R_{\in}(x, z)$ iff there is y amongst the yy such that $R_{\in}(x, y)$. These are all the mathematically important properties of the "open sets": that they are literally collections of things falling under P is arguably not. There are ways one could require the things falling under P_{open} to be more like collections, for instance taking them to be fusions of things falling under P (if the latter are points, or atoms in some sense), or otherwise properties or pluralities – if one accepts some form of higher order logic in which one can take collections of these in turn, so that one can state the property of being closed under unions.

We can then characterize variants of the notion of topological space, such as that of ringed space – a topological space equipped with a sheaf of rings on it. A sheaf of rings here is a presheaf of rings that satisfies two extra conditions, where a presheaf of rings is an assignment of a ring $\mathcal{O}_X(U)$ to each open subset U of X, and a ring homomorphism $r_{V,U} : \mathcal{O}_X(U) \to \mathcal{O}_X(V)$ for each pair V, U of open subsets with $V \subseteq U$. The two extra conditions on a sheaf are firstly, that if $(U_i)_{i \in I}$ is a cover of U (i.e. we have $U = \bigcup_{i \in I} U_i$), and we have $a, b \in \mathcal{O}_X(U)$ with $r_{U_i,U}(a) = r_{U_i,U}(b)$ for all i then $a = b$; and secondly, that if $(U_i)_{i \in I}$ is a cover of U and we have a family $(a_i)_{i \in I}$ where $a_i \in \mathcal{O}_X(U_i)$ for all i, such that $r_{U_i \cap U_j, U_i}(a_i) = r_{U_i \cap U_j, U_j}(a_j)$ for all i, j, then there is $a \in \mathcal{O}_X(U)$ such that $r_{U_i,U}(a) = a_i$ for all i. To characterize this notion we can take the above characterization of a topological space in terms of predicates P and P_{open} and relation $R_{\in}$, and introduce a new binary relation $R_{\mathcal{O}}$, ternary relations R_+ and $R_\times$ and a four place relation R_r. We

use $R_{\mathcal{O}}(a, U)$ to play the part of the relation of a being an element of $\mathcal{O}_X(U)$, requiring that this only holds of a and U if U is open, and requiring (for convenience) that if $U \neq V$ then we don't have both $R_{\mathcal{O}}(a, U)$ and $R_{\mathcal{O}}(a, V)$, in other words that the sets $\mathcal{O}_X(U)$, $\mathcal{O}_X(V)$ are disjoint. Then we use R_+ and $R_\times$ to play the role of addition and multiplication in every $\mathcal{O}_X(U)$, stating that if there is U such that $R_{\mathcal{O}}(a, U)$ and $R_{\mathcal{O}}(b, U)$ then there is a unique c such that $R_{\mathcal{O}}(c, U)$ and $R_+(a, b, c)$, and a unique d such that $R_{\mathcal{O}}(c, U)$ and $R_\times(a, b, d)$; and also stating that if $R_+(a, b, c)$ then there is some U such that $R_{\mathcal{O}}(a, U)$, $R_{\mathcal{O}}(b, U)$ and $R_{\mathcal{O}}(c, U)$, and stating the equivalent for $R_\times$. We state that these operations satisfy the usual ring axioms, in effect giving us a ring structure on each $\mathcal{O}_X(U)$. Then we state that for each open set U and V, if $V \subseteq U$ then for each a with $R_{\mathcal{O}}(a, U)$ there is a unique b with $R_{\mathcal{O}}(b, U)$ and such that $R_r(V, U, a, b)$, and moreover that whenever $R_r(V, U, a, b)$ then U and V are open and $V \subseteq U$ and $R_{\mathcal{O}}(a, U)$ and $R_{\mathcal{O}}(b, U)$. Thus this relation $R_r(V, U, a, b)$ represents the relation of $r_{V,U}(a) = b$ holding, and we can state that it is a ring homomorphism. This so far gives us a characterization of the structure of a presheaf of rings on our topological space, and stating the two further conditions that make it a sheaf is not difficult, as we can use pluralities UU of open sets in place of families $(U_i)_{i \in I}$ when discussing covers, and similarly in the second condition instead of discussing a family $(a_i)_{i \in I}$ where $a_i \in \mathcal{O}_X(U_i)$ for all i, discussing a plurality aa where for each U amongst our plurality UU there is a unique a amongst aa such that $R_{\mathcal{O}}(a, U)$ (for this to work, we use the fact that there are no a satisfying $R_{\mathcal{O}}(a, U)$ and $R_{\mathcal{O}}(a, V)$ with $U \neq V$). Putting this all together, we have a faithful self contained real world characterization of the notion of ringed space. One could then try to build on it to obtain the important mathematical notion of scheme, a particular kind of ringed space, though that concept brings additional complexities.

Finally we consider the concept of real Banach space – a normed real vector space which is complete. This case is unlike most of the previous ones, in that the natural way to characterize the structure does not give a self contained characterization – we

want auxiliaries in the form of a complete ordered field structure, to provide the scalars for scalar multiplication of elements of the vector space (the issue of self contained vs auxiliary involving characterization, and the example of real vector spaces, were discussed above). Apart from this feature, characterizing the notion of normed real vector space is straightforward. A further subtlety arises though when we consider the property of completeness. This is normally defined by quantifying over sequences taking values in the space, and stating that all such sequences which are Cauchy converge to some point in the space. A sequence here is a function from $\mathbb{N}$ to the space; we have a copy of $\mathbb{N}$ available, lying in our complete ordered field, but functions themselves will be a further posit, and another step away from a self contained axiomatization. However we can actually avoid the need for this further posit, by judiciously modifying the definition of completeness. Instead of phrasing this in terms of Cauchy sequences, we can introduce the (set theoretic) concept of a *Cauchy subset* of a metric space X, which is a subset Y such that for every $\epsilon > 0$ there is a finite subset Y' of Y such that if $x, x' \in Y \setminus Y'$ then $d(x, x') < \epsilon$. It is not difficult to see that every Cauchy subset is finite or countably infinite,[10] that if $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence then its image $\{x_n \mid n \in \mathbb{N}\}$ is a Cauchy subset, and that any Cauchy subset is the image of a Cauchy sequence. Moreover we can define what it is for a Cauchy subset Y of X to converge to a point $a \in X$, namely that for all $\epsilon > 0$ there is a finite subset Y' of Y such that if $x \in Y \setminus Y'$ then $d(x, a) < \epsilon$; and it is easy to see then that if a Cauchy sequence has infinite image then it converges to $a \in X$ iff its image does (it is a feature of this definition of convergence for Cauchy sets that if finite then they converge to every point in the space). It then follows quickly that the statement that a metric space X is complete is equivalent to the statement that any Cauchy subset of X converges to some point in X. We can then take this as a definition of completeness – though not standard, it is perfectly valid – and finish our

[10] If Y is a Cauchy subset then for each $n \in \mathbb{N}$ we can find a finite subset Y_n of Y such that if $x, x' \in Y \setminus Y_n$ then $d(x, x') < \frac{1}{n}$. Then if $y, y' \in Y \setminus \bigcup_n Y_n$ then for all n we have $d(y, y') < \frac{1}{n}$ and so $y = y'$. Thus $Y = (\bigcup_n Y_n) \cup (Y \setminus \bigcup_n Y_n)$ is a countable union of finite or countable sets, so is finite or countably infinite.

characterization of Banach space by stating its equivalent in plural, namely logic that any "Cauchy plurality" (paralleling the definition of Cauchy set) converges to some point in the space. Thus we can avoid having to use a collection of sequences as auxiliary objects, by reconsidering the set theoretic definition we are seeking a faithful characterization of. One can of course characterize real Hilbert spaces in a similar way, and complex Banach and Hilbert spaces given a characterization of the complex numbers.

We won't discuss in detail how these structures might be physically realised – though Hilbert spaces are an essential tool in areas of physics like quantum mechanics and quantum field theory – but as a general point, it has often been noted how concepts originally defined and investigated out of purely mathematical interest regularly end up later playing a crucial role in real world applications of mathematics (see for instance Rowlett 2011). The uses of Hilbert spaces in quantum mechanics and smooth manifolds in general relativity are two standard examples, both concepts being originally developed with no awareness of their possible applications in physics. The range of mathematical structures used in physical theories is very great, and we have no way of telling which kinds of mathematical structures will ultimately end up playing a role in them – let alone which structures might actually find some manner of physical instantiation. After all fundamental physics is still, it appears, a decidedly unfinished endeavour, with many difficult questions remaining open – how should quantum mechanics be reconciled with general relativity? Is spacetime fundamentally continuous or discrete? Is there a more fundamental level of explanation than the standard model, which can explain apparently ad hoc features such as the values the (two dozen or so) physical constants take? What is the status of dark matter and energy? And so on. There may be a worry like that seen above expressed by Maddy (2011, pp. 27–28), who considers the applications of mathematics we tend to make currently and concludes that we are rarely able to establish anything like an isomorphism between the physical realm and a complex mathematical structure; often the relationship, she argues, between the mathematical and the physical

is much less straightforward. However one can just as easily put a different interpretation on the situation, where the fact that physics makes essential use of serious mathematics but without confirming its literal truth in the world is not taken as evidence that only simple mathematical structures are physically instantiated – but to the contrary, taken as a sign of how complex the mathematical structure of the world must really be.

Now in all these cases we have been seeking real world characterizations of set theoretic concepts. It should be pointed out that in each case one can generally obtain characterizations of "the same" concept in other proof systems as well, such as ETCS and homotopy type theory. This was discussed above for the case of groups, but applies much more widely. In ETCS, one can use what is called Mitchell-Bénabou language to interpret set theoretic formulae with bounded quantifiers – quantifiers of the form $\forall x \in y$ and $\exists x \in y$, where y is a set, instead of $\forall x$ and $\exists x$ – in the context of the category of sets (see for instance MacLane and Moerdijk 1994, §VI.5). One uses arrows $X \to \Omega$, where Ω is the subobject classifier (representing the set of truth values), to play the role of propositions, with categorical operations giving conjunction, disjunction, implication, negation, and bounded quantification. One has (categorical) power sets available in this context, and can use for instance subobjects of $X \times Y$ to play the role of a relation from X to Y. Thus all standard set theoretic definitions of structures can be carried out in ETCS in much the same way: indeed one of the claims advocates of ETCS make is that high level informal proofs according to the principles of ETCS are effectively indistinguishable from high level informal proofs according to the principles of ZFC.[11] Finding real world characterizations of these kinds of set theoretic definitions in ETCS is then much the same challenge as finding real world characterizations of set theoretic definitions in ZFC. There are ways concepts can be defined categorically which don't have such an immediate analogue in ZFC, such as the definition of a natural numbers

[11] As mentioned in section VI.2, the main difference is that separation in ETCS only applies to formulae with bounded quantifiers, and there is no replacement scheme. One can extend the theory to address these issues though.

object in terms of the ability (essentially) to define morphisms by recursion out of it (ibid., §VI.1), but in as much as such a formal definition gives "the same" concept as the usual definition in ZFC, it thus also gives "the same" concept as that of simply infinite sequence as defined above.

In homotopy type theory, similar remarks apply. The correspondence between propositions and types allows one to use define kinds of structures in terms of types, including definitions in higher order logic, using the presence of universes to allow one to quantify over "all properties" in a sense, though with the subtlety that such a definition is relative to a choice of universe (Univalent Foundations Program 2013, §1.11). Definitions of this kind may then be employed to parallel set theoretic definitions like those considered here. In the theory one can also define a "cumulative hierarchy type" as a higher inductive type, which models the set theoretic hierarchy, and if one adds the axiom of choice to the theory then this hierarchy satisfies the analogues of all the axioms of ZFC, allowing one to directly carry out set theoretic definitions. When definitions of structures in homotopy type theory parallel those in ZFC, obtaining real world characterizations of them is essentially the same problem. Again there are other kinds of definitions in homotopy type theory that do not have direct set theoretic analogues (such as definitions involving the unique properties of identity type) – but if such a definition gives "the same" concept as a definition in ZFC, then a real world characterization of the ZFC concept will also be a real world characterization of the homotopy type theory concept.

Before moving on, there are a few final issues to discuss regarding the subject of this section. One is that though the focus here has been on potential physical realizations of mathematical structures, there is no requirement that realizations be physical in any sense. When we characterize a structure in the manner discussed in this section, we state a characterization which can be satisfied or fail to be satisfied by real world objects, by objects that really exist – whatever we mean by these kinds of terms. If one believes that there might be real objects which exist, but exist not in space or

time – what are generally called abstract objects – then there is nothing to stop such abstract objects from forming realizations of any of the concepts discussed here (or other mathematical concepts). There could for instance be a simply infinite sequence of abstract objects, either one which just "happens" to exist in some sense, or one made up of distinguished objects with special properties. *Ante rem* structuralists and Neo-Fregeans both defend a version of the latter view with regard to the simply infinite sequence $\mathbb{N}$ of natural numbers. There could also be abstract complete ordered fields, groups, graphs, topological spaces, manifolds, schemes, and so on and so forth. I am not going to argue for the existence of the abstract objects necessary to form such structures, but it is worth noting that (as far as I'm aware) arguments *against* the existence of such abstract objects have rarely been put forward. Nominalism – the view that abstract objects do not exist – is not uncommon, but arguments for nominalism generally stem from worries about how we could know about abstract objects, or refer to them – thus casting doubt on the advisedness of relying on abstract objects – rather than directly arguing that abstract objects do not exist. In fact these kinds of nominalist arguments can even be taken as telling against confidence that abstract objects do not exist, since how could knowledge that they do not exist be any easier to come by than knowledge that they do exist? Thus even if a belief in abstract objects may be hard to fully justify, a belief that they categorically do not exist seems to be equally unwarranted – so that the possibility of mathematical structures formed out of abstract objects appears to be a live one.

We should consider here, at least briefly, the status of objects that exist on a deductivist understanding of mathematics. For instance one could advocate a deductivist understanding of ZFC, along the lines of Tait (2005) and Muller (2004), in which the sense of set theoretic statements is given by the rules for asserting or denying them according to the axioms of ZFC and the principles of first order logic.[12] Thus one can

[12]Tait actually advocates a type theoretic understanding of logic, but that is inessential here.

rightly assert the existence for instance of a set like $\{\varnothing, \{\varnothing\}, \{\{\varnothing\}\}\}$, since it can be proved to exist by the reasoning available. Should the objects which "exist" in this sense be taken to "really exist" – to be included in the range of things we talk about when we say "there are" and "every thing" in English? I believe Tait and Muller would argue that they should, though they do not (as far as I can see) explicitly address the question. If they are right – and I will remain agnostic on this – then structures formed from such objects will be examples of real world instantiations of mathematical concepts. A further subtlety arises with structures that we characterize in plural logic, for instance the notion of simply infinite sequence or complete ordered field. One might assume that if we have a set equipped with a successor function which satisfies the set theoretic definition of simply infinite sequence, then it will also satisfy the plural logic definition; but this does not follow. When one has the kind of deductivist understanding of set theory that Tait and Muller advocate, the separation scheme – for forming subsets – is limited to properties definable in set theoretic language, and thus set theoretic simply infinite sequences can only be argued to satisfy induction for properties definable in this language. There is thus no way to argue that such a set theoretic simply infinite sequence satisfies the full plural logic inductive clause, as the language of plural logic is not allowed into the set theoretic inductive clause; so there is no way of arguing that a set theoretic simply infinite sequence, on the deductivist view, is a simply infinite sequence as defined here. The same goes for complete ordered fields. This may seem like a minor technical point, but actually reflects a deep mismatch between mathematical concepts as defined within a deductivist setting, and mathematical concepts as faithfully characterized in a logic properly capturing their informal content. Because deductivists advocate an understanding of mathematics based on a closed system of precise rules governing the mathematical vocabulary,[13] the mathematical vocabulary ends up semantically isolated from real world vocabulary, and a mathematical principle like that of induction

[13]Tait has an open ended conception, with us able to add new rules and axioms to our proof system, but at any given time we have only some fixed system in mind.

which should hold of "all properties" or "all pluralities" or "all subsets" ends up in fact only applying to those properties definable in the mathematical theory – rather than all properties the full power of our natural language is able to define. This is related to the point that we can have a model M of ZFC in which the set of natural numbers ω is actually a non standard model of arithmetic, satisfying the induction axiom in the model:

$$M \vDash \forall x \ (x \subseteq \omega \text{ and } \varnothing \in x \text{ and } \forall y \ (y \in x \rightarrow y \cup \{y\} \in x)) \rightarrow x = \omega$$

but where there is in reality an element a such that $M, v(x \mapsto a) \vDash a \in \omega$ and $M, v(x \mapsto a) \vDash a > \underline{n}$ for every numeral $\underline{n}$ (where v is a variable assignment). In this sense though the definition of "simply infinite sequence" in set theory and in plural logic above both aim to capture the same informal concept, they need not be extensionally equivalent – we can see what the intention behind the set theoretic definition is, but the definition may still not be a fully adequate one.

One can make clearer what is lacking from the deductivist understanding of set theory by considering an alternative view, in which one obtains justification for an open ended understanding of the set theoretic axiom schemes (rather than them just being limited to mathematical vocabulary). For instance Boolos (1971) discusses the iterative conception of sets, in which we start with the set of all things which exist which aren't sets, and form every set all of whose elements are amongst these objects, and then form every set all of whose elements are amongst these objects (including the sets formed), and keep going, iterating this into the transfinite. At each stage we are forming every possible set, whether such a subset is definable or not, and whatever language such a subset may or may not be definable in. Thus we can see (if we believe this conception) that when using the separation axiom scheme to form the subset $\{x \in y \mid \phi\}$ of a set y, we can use any vocabulary whatsoever in defining the formula ϕ, not just the vocabulary

of set theory (with a possible restriction to vocabulary such that the formula ϕ defines a sharp property). Thus we obtain what we call an open ended separation scheme – expanding its range as we add new vocabulary to our language. Given this, the above problem with the deductivist understanding of set theory is no longer an issue. Indeed we have a set theoretic simply infinite sequence (x, r_s), with r_s its successor function, then defining the relation R_S by $R_S(a, b)$ iff $r_s(a) = b$ we obtain that R_S is a successor function in the sense of plural logic, defined above. Here we use in crucial fashion the ability to use all vocabulary of our language, including plural variables, in the separation scheme. Similarly from a successor function R_S as defined above in plural logic, we can obtain a set theoretic infinite sequence.[14] The same goes for complete ordered fields.

If one does regard set theoretic language as genuinely meaningful language in this way, then one can regard characterizations of structures in set theory as real world characterizations – with no need to reinterpret them in the logics used here. As just seen, if we have an open ended conception of the set theoretic axiom schemes, we can often obtain equivalences between set theoretic definitions and the kinds of definitions given above in restricted predicative second order logic and plural logic. When working mathematically with sets, they are often conceived of as pure sets – with all sets only having other sets as their members, rather than other real world objects (like the laptop and chair in the initial group example). One would have to phrase one's set theory as an impure set theory, allowing sets of urelements, for set theoretic characterizations of structures to allow for structures containing objects that aren't sets. Apart from this the main difference between set theoretic definitions of structures and the kinds of definitions given above is that set theoretic structures are generally taken to be sets, rather than proper classes, whereas for instance a group as defined above could consist of a group multiplication operation defined on a proper class of elements, such as the

[14]We can use here the double ancestral to define a primitive recursive function from the finite ordinals to the domain of R_S, thus obtaining by open ended replacement that the domain of R_S is a set. The rest is easy.

surreal numbers. If one has the kind of understanding of set theory discussed here, then one could make the case that a faithful characterization of a concept of group should limit itself to set sized structures, rather than proper class sized structures. As with other issues, I will remain agnostic on this.

5 The eliminative constraint

We have seen how though mathematical concepts are often given official definition as formal concepts in the context of a mathematical proof system, we can generally find real world characterizations which capture the same informal concept as the formal definition. This was discussed primarily for the case of ZFC in section VI.4, with the cases of ETCS and homotopy type theory also being touched on.

Now we use this to argue for a constraint we should put on interpretations of mathematics, specifically interpretations of what generalizations about mathematical structures mean. This will apply to generalizations about all kind of mathematical structures that we can obtain real world characterizations of, including the examples – groups, rings, topological spaces, Banach spaces and so on – seen in section VI.4, and we will also discuss the case of statements about the distinguished structures $\mathbb{N}$ and $\mathbb{R}$.[15]

By generalizations about a kind of structure we mean statements in English, informally universally quantifying over structures of that kind and stating some property of them. A generalization about groups, for instance, is just a statement ϕ of the form

$$\text{For every group } G, G \text{ satisfies } \mathcal{Q}(G)$$

where $\mathcal{Q}(G)$ is some property that any group G does or doesn't satisfy – for instance $\mathcal{Q}$ might be the property G has a unique identity element, or that for $g \in G$, if there is

[15]Philosophical interpretations of mathematics have typically focused on what mathematical generalizations about kinds of structures, or about particular structures such as $\mathbb{N}$ and $\mathbb{R}$, mean. Later in this section it will be discussed how though many mathematical results are of this form, there are also many which are not, and the question of interpreting the latter will be considered.

$h \in G$ with $gh = h$ then g is the identity element of G. As discussed in section VI.3, a variety of interpretations $I : \phi \mapsto \phi^I$ of such mathematical generalizations have been put forward. In addition, as discussed in section VI.4, we can characterize as a formula $\text{Group}(P, R_\times)$ what it is for a predicate P and ternary relation $R_\times$ to "be" a group, capturing the same informal concept of group as that defined for instance in ZFC. Then if $\mathcal{Q}$ is a property stated informally in English that a group G can satisfy or fail to satisfy, we call $\mathcal{Q}$ *definable* if it can be stated using our logical resources as a property $\mathcal{Q}(P, R_\times)$ purely in terms of such a predicate P and relation $R_\times$ and the objects which it relates, without requiring other auxiliary objects.[16] Whether a property is definable can depend on what logical resources we take there to be available: the property of G having a unique identity element is straightforwardly definable (provided we accept first order logic), but the property that "if G is finitely generated and abelian then every subgroup is abelian" requires a stronger logic, for instance ancestral plural logic as described in chapter V.

The latter example brings up another incidental issue not related to definability, which is that instead of thinking in terms of a generalization

For every group G, G satisfies $\mathcal{Q}(G)$

where $\mathcal{Q}(G)$ is this complicated property, it may be more natural to think in terms of a generalization

For every finitely generated abelian group G, G satisfies $\mathcal{Q}'(G)$

where $\mathcal{Q}'(G)$ is the property of G having only finitely generated subgroups. One can then give a faithful real world characterization in ancestral plural logic of the concept of what it is for a predicate P and ternary relation $R_\times$ to define a finitely generated abelian group, and the property of having only finitely generated subgroups is then a definable

[16] As with the notion of a faithful real world characterization of a formal concept, this notion of definability is not a precise one, and whether a definition of an informally stated property accurately captures it may be open to debate.

property of such a structure. Nothing said here depends on which of these options is taken.

Now suppose we are given an interpretation $I : \phi \mapsto \phi^I$ of some class of mathematical statements, which includes the generalizations about groups. The eliminative constraint on I is the condition that for all such generalizations ϕ, if ϕ is of the form

$$\text{For every group } G, G \text{ satisfies } \mathcal{Q}(G)$$

where $\mathcal{Q}$ is some definable property of groups, then if ϕ is interpreted by I as true – if ϕ^I holds – then it must be the case that

$$\text{For every } P, R_\times, \text{ if } \text{Group}(P, R_\times) \text{ then } \mathcal{Q}(P, R_\times)$$

holds. In other words, whenever the statement that all groups have property $\mathcal{Q}$ is interpreted as true, it must be the case that all realizations of the concept of group do actually have property $\mathcal{Q}$.

Of course, this constraint is not intended to just apply to generalizations about groups. In general, if S is a kind of structure of which we can obtain a faithful real world characterization, and $\mathcal{Q}$ is a definable property of S's, then we call the statement "all S's satisfy $\mathcal{Q}$" a definable generalization (about S's). Suppose that $I : \phi \mapsto \phi^I$ is an interpretation of some class of mathematical statements, which includes generalizations about some kind of structure S of which we can obtain a faithful real world characterization. Then the eliminative constraint on I requires that for all generalizations ϕ stating that a definable property $\mathcal{Q}$ holds of all S's, if ϕ is interpreted by I as true then every realization of S satisfies $\mathcal{Q}$. The latter statement, that every realization of S satisfies $\mathcal{Q}$, is just the eliminative structuralist interpretation ϕ^{Elim} of ϕ. Thus we can state the eliminative constraint more briefly as the condition that for all definable generalizations ϕ interpreted by I, ϕ^I implies ϕ^{Elim}. Hence the term "eliminative constraint".[17]

[17]Incidentally, it should be mentioned that rather than just thinking in terms of kinds of structures being the standard examples such as groups, topological space, schemes and so on, one can think of

The reasoning behind this constraint is simple. The purpose of a faithful real world characterization of a kind of mathematical structure S is to characterize "the same" concept, but in a manner which can be satisfied or fail to be satisfied by real world objects. To assert that every ring has property $\mathcal{Q}$ while accepting the existence of a realization of the concept of ring which does not have property $\mathcal{Q}$ could only be coherent if our real world characterization of "ring" differed in some significant sense from our pre-existing conception of what a ring is – but this is exactly what our characterization being faithful rules out.

However it could, perhaps, be argued that such generalizations just do have an agreed upon hidden meaning, which accepts this kind of clash as a possibility. There is no evidence of this in everyday discussion of mathematical statements by mathematicians, however: if a mathematician sat through a seminar concerning a property $\mathcal{Q}$ of topological spaces, and at the end said they agreed it was true that every topological space satisfies $\mathcal{Q}$, but nevertheless there could be a real example of a topological space for which $\neg\mathcal{Q}$ held, they would be met with bafflement. Mathematicians do sometimes advocate consequentialist or if-then interpretations of mathematics when musing philosophically (rather than when actually doing mathematical research), but never with awareness that this could lead to such a clash between the apparent meaning of mathematical generalizations and their actual import. Moreover it is only the apparent meaning of mathematical generalizations that makes sense of the way we regard them, celebrating their proofs as informing us of a fact about the world, and relying on them whenever relevant in physics or other sciences; if one proves the result that all groups satisfy $\mathcal{Q}$, but this is compatible with some real world group still having property $\neg\mathcal{Q}$, then why should the result license regarding all groups used in applications of mathematics

adjective-noun combinations such as "compact Hausdorff topological space", "torsion free abelian group" and the like as naming kinds of structures, and apply this constraint to these more general kinds of structures. This is in line with the observation from above that the statement that for every group, if it is finitely generated and abelian then its every subgroup is finitely generated, is perhaps best viewed as a generalization about finitely generated abelian groups rather than about groups.

– whether literal or idealized – as satisfying $\mathcal{Q}$?

Note that we do not require here that we actually settle on eliminative structuralism as the correct interpretation of mathematical generalizations. It is perfectly reasonable to advocate some other interpretation – just as long as it is compatible with eliminative structuralism in the above sense. For instance one might fix some mathematical proof system T, and use a consequentialist interpretation with respect to T in which a generalization ϕ about a kind of mathematical structure S is interpreted as meaning that $T \vdash \phi^T$, where ϕ^T is a reduction of ϕ to the language of T via a formal definition in T of what a structure of kind S is. There is nothing incoherent about appealing to formal definitions in a proof system like this instead of the kinds of real world characterizations seen in section VI.4: they can perfectly well characterize "the same" concept, and can be more convenient, as a proof system may be a more expressive setting for defining concepts – with a full range of auxiliary objects available – than the logics used in section VI.4, where sometimes fairly subtle manoeuvres were needed. We just require that the resulting interpretation doesn't mislead us as to the properties of realizations of structures that actually exist.

The eliminative constraint as stated above applies to the concepts of simply infinite sequence and complete ordered field as well as to the other examples discussed. However though these are key concepts in mathematics, generalizations about them are rarely stated – for the simple reason that all simply infinite sequences are isomorphic, as are all complete ordered fields, so that any arithmetic fact that holds of one infinite sequence holds of all of them, and similarly for any fact about complete ordered fields stated in the language of ordered fields. Thus it is common to state one's result as simply about the paradigmatic simply infinite sequence $\mathbb{N}$, or the paradigmatic complete ordered field $\mathbb{R}$, with it being understood that any such result will hold of all structures of the relevant kind. Since these distinguished structures are intended as paradigms in this sense (and statements about them in proof systems are equivalent to generalizations about the

relevant structures) the eliminative constraint applies to statements about them just as much as to other explicit generalizations: if we have an interpretation I of some class of statements which includes statements of arithmetic about $\mathbb{N}$, then for I to be acceptable we require that whenever ϕ states that some arithmetic property $\mathcal{Q}$ holds of $\mathbb{N}$, if ϕ is interpreted as true by I then every realization of the concept of simply infinite sequence satisfies $\mathcal{Q}$. In other words we require again that ϕ^I implies ϕ^{Elim}. The equivalent condition is required for statements in the language of ordered fields about $\mathbb{R}$.

As with other generalizations, ϕ^{Elim} is required to hold of all realizations that exist, not just those we know about. The existence of a real example of a topological space satisfying $\neg\mathcal{Q}$ is just as incompatible with the assertion "all topological spaces satisfy $\mathcal{Q}$" whether we know about it or not. As discussed in section VI.4, the existence of physical realizations of many kinds of mathematical structures is a live possibility that cannot currently be ruled out; so – even though we may not yet have confirmed the existence of such realizations – one cannot currently rely on the eliminative constraint being vacuous in these cases.

The eliminative constraint has been stated here only in terms of the actual realizations that exist. However if one takes talk of possibility and necessity seriously, I think there is a case that the eliminative constraint should apply not just to all actual realizations, but to all possible realizations: a generalization about a kind of structure being true should not rely on there only being a limited range of realizations of that structure in our world as it happens to be, but should hold no matter what realizations of that structure could exist. In this case one obtains a stronger constraint, requiring that if ϕ is a definable generalization interpreted as true then ϕ^{Elim} doesn't just hold accidentally, but holds necessarily. In other words we require that ϕ^I implies ϕ^{Modal}. One could call this stronger constraint the modal constraint.

The eliminative constraint as discussed above is only applied to mathematical generalizations about a single kind of structure. Of course, many mathematical statements

are not of this restricted form. Some mathematical statements are effectively universal generalizations about multiple structures – for instance statements about the properties of certain kinds of homomorphism or other maps between structures of some kind. In such instances there is a case again for a form of the eliminative constraint, where we require that ϕ^I implies ϕ^{Elim}, where this time ϕ^{Elim} quantifies over multiple realizations of the relevant kind of structure, or maybe over realizations of different kinds of structures, perhaps with some relationship between them. For example suppose ϕ states that a ring homomorphism is injective iff its kernel is $\{0\}$; this version of the eliminative constraint would state that for ϕ to be interpreted as true, whenever we have two realizations of the concept of ring (quantifying over instantiations of this concept twice), and a suitable relation R_f between them which defines a homomorphism, we need that this homomorphism is injective iff the only object mapped to 0 in the codomain is 0 in the domain. One could just think of such examples as a case of a single larger joint structure.

Then naturally there are existential statements, or statements of the logical form $\forall\exists$ – for instance the statement that a complete ordered field exists, or that any topological space has a compactification, or a fundamental group and so on. One might be tempted to argue for some sort of eliminative constraint on the interpretations of such statements, but I think the case here is a much less strong one. The "generalization" form of the eliminative constraint requires that mathematical generalizations do not mislead us about the properties of realizations of structures that actually exist; and this, I think, is just a greater defect than stating the existence of structures which are not actually instantiated. The latter is an accepted aspect of mathematical discourse – it is well known that the existence of the structures mathematics describes is a difficult and controversial question, both amongst mathematicians and philosophers. When mathematics justifies the existence of certain structures, science may well make use of them in models and theories, but what the proper scientific attitude to the existence of such structures should be is a difficult question; there are no such reservations concerning mathematical

generalizations in applications, where any generalizations mathematics takes to be true of a certain kind of structure will happily be applied, if relevant, to instances of that structure featuring in models or encountered in reality. Thus the eliminative constraint will not be applied here to existential statements about structures – though of course existential statements within structures, such as the statement that in a simply infinite sequence there are infinitely many primes, can count as generalizations about structures and so come under its scope.

What counts as an existential statement "within" a structure as opposed to an existential statement about a structure may depend on what kinds structures we are considering – for instance, any existential mathematical statement could potentially be thought of as an existential statement within a broader mathematical universe such as the set theoretic hierarchy. One can of course apply the eliminative constraint to such statements about a set theoretic hierarchy – it would require that if a generalization ϕ about set theoretic hierarchies (or perhaps about *the* set theoretic hierarchy, regarded as a paradigm) is interpreted as true then ϕ^{Elim} needs to hold, where ϕ^{Elim} restates ϕ as a generalization about realizations of the concept of set theoretic hierarchy, via some suitable faithful real world characterization (for instance using plural logic). Of course it is implausible that there could be a *physical* realization of this concept, as its cardinality would be far too large, but – as discussed in section VI.4 – there do not appear to be good reasons to rule out the existence of abstract objects, so that the possibility of some domain of really existing abstract objects having the structure of a set theoretic hierarchy is a live one. If one believes in the iterative conception of set, then that would provide an example of such a domain.

Now we connect this discussion of the eliminative constraint with the discussion of the notion of genuine proof from section VI.2. As seen there, the notion of genuine proof is one in which proof is factive – for a statement ϕ to be provable, it must be the case that ϕ is actually true. To genuinely prove a statement we thus need some interpretation

of what it means (so that it is the kind of thing that can be true or false). As discussed here though, we should reject any interpretation that does not satisfy the eliminative constraint. Thus, being able to genuinely prove ϕ requires that ϕ^I holds for some I where ϕ^I implies ϕ^{Elim}, and thus requires that ϕ^{Elim} holds.

We do not need to be over eager with the condition that we require an interpretation of ϕ in order to be able to genuinely prove it. We need an interpretation that doesn't immediately collapse under scrutiny, but do not necessarily need to have a fully developed philosophical account – in the same way as to argue whether something is frivolous, you need to know what "frivolous" means, but do not need a philosophical theory of frivolity. Also, it is perfectly possible for multiple people to all regard an argument p as a genuine proof of ϕ, whilst each having their own differing interpretation of what ϕ means – for instance all regarding a proof in first order arithmetic as a proof of an arithmetic statement, whilst disagreeing about what the proper interpretation of the result is. Thus to be able to regard some argument as a genuine proof of ϕ (where ϕ is of the form discussed above), it just needs to be the case that there is some available interpretation of ϕ which interprets it as true and satisfies the eliminative constraint – we do not necessarily have to agree on any such particular interpretation, or have a fully realized philosophical theory of it.

The standard worry about eliminative structuralism is the threat of vacuity: that there may not actually be any structures of the given kind. This possibility may in fact weaken the eliminative constraint, but it is only if we know that it is vacuous for some particular kind of structure that it becomes relevant. To regard an argument p establishing ϕ as a genuine proof, one must be able to *know* that ϕ is true on acceptable interpretation – if it is merely probable that it is true on acceptable interpretation, the argument is not a proof – and thus requires knowing that ϕ^{Elim} holds. If the eliminative constraint for generalizations like ϕ is in fact vacuous, but we don't know this to be the case, then this doesn't allow us to regard an argument p establishing ϕ as a proof.

To end the section, we discuss which of the interpretations seen in section VI.3 satisfy this eliminative constraint. Firstly, the consequentialist, deductivist, and if-thenist interpretations. For consequentialism and if-thenism, here the distinction between local and global versions is again key. If for instance we have a local consequentialist interpretation of a statement ϕ of arithmetic, interpreting ϕ as meaning that for instance $\mathrm{PA} \vdash \phi^{\mathrm{PA}}$, then this trivially satisfies the eliminative constraint: if ϕ is provable in PA, then it will hold in any simply infinite sequence, as all the axioms of PA hold in any simply infinite sequence (and deductions in first order logic preserve logical entailment). On the other hand if we take a global consequentialist interpretation, interpreting ϕ as meaning that $\mathrm{ZFC} \vdash \phi^{\mathrm{ZFC}}$, then it is not at all clear whether ϕ being "true" will imply that it holds in any simply infinite sequence. This is an example of the question of soundness for ZFC, discussed in section VI.6. The same goes for statements about groups, and topological spaces, and other kinds of structure: if one takes a local consequentialist interpretation, with regard to an axiomatization of the structure of interest, then the eliminative constraint may be trivially satisfied; but for global consequentialism, with regard to a general mathematical axiom system, it is an example of a the soundness question discussed in section VI.6, and may be a question requiring substantial thought. The outcomes for if-thenism are similar. The eliminative constraint for deductivism is similar to that for global consequentialism – a statement of arithmetic ϕ being assertible amounts to ϕ being provable according to the rules of our chosen proof system T, and so the eliminative constraint amounts to the question of whether provability in T implies that an arithmetic statement actually holds of all infinite sequences, which is the question of (arithmetic) soundness for T.

As noted above, the eliminative structuralist interpretation trivially satisfies the eliminative constraint. Modal structuralism also trivially satisfies the eliminative constraint, moreover actually trivially satisfying the stronger modal constraint. *Ante rem* structuralism satisfies the eliminative constraint as well, given the posit that for any

realization of a kind of structure, the abstract form of the isomorphism type of that realization exists; the *ante rem* structuralist interpretation ϕ^{ante} of a statement ϕ then implies the eliminative structuralist interpretation, since any realization of a kind of structure is isomorphic to the form of its isomorphism type. For the cases of complete ordered fields and simply infinite sequences, we use here the relevant categoricity results. The Neo-Fregean interpretation of arithmetic satisfies the eliminative constraint, again because of categoricity for simply infinite sequences.

If one interprets the conditional in the statement of the eliminative constraint as a material conditional, then the fictionalist interpretation actually trivially satisfies the eliminative constraint, as ϕ^{fiction} is always false, so that $\phi^{\text{fiction}} \to \phi^{\text{Elim}}$ is always true. However I think this example just suggests that the implication in the eliminative constraint should be understood as a natural language if-then, rather than a material condition, in which case the status of fictionalism becomes unclear. It is possible though that a fictionalist could accept a form of the eliminative constraint as a condition on axiom systems; this becomes an example of the requirement of soundness, as discussed in the next section.

Finally there is set realism. How this stands with regard to the eliminative constraint is an interesting and important question, that will be returned to in section VI.7.

6 Foundational goals

Now we turn to the central questions of this chapter: what the purpose of a foundation for mathematics is, and what we should look for when deciding on one.

When looking for a foundation for mathematics we are choosing between proof systems – for instance ZFC, ETCS, homotopy type theory, PA^2, Quine's theory New Foundations, amongst many others. As discussed in section VI.2, any such proof system T comes equipped with a precise notion of what it is for a string of symbols to be a

formal proof (or formal derivation) in T. If the workings of this notion of formal proof are suitably amenable to being grasped by humans, then it can give rise to a corresponding notion of informal proof in T, consisting of informal arguments that take place at a range of levels of detail/compression, from those which are highly detailed and close to the level of formal proof in T, all the way to those which are very highly compressed, for which formalization may be a lengthy process, though still always possible in principle if a proof is valid. This was discussed for the case of informal proof in ZFC in chapter I.

Not just any proof system can serve the role of a foundation, however. Those proof systems that are considered candidate foundations are ones in which a wide range of mathematical structures can be represented and investigated. This was discussed for the case of ZFC, ETCS and homotopy type theory in section VI.4, where it was seen that sets and types can be used both to characterize and construct all manner of structures. The same goes for Quine's theory New Foundations, and PA^2 also has some potential in this regard, with a range of structures being representable in it via coding.

Such proof systems can thus provide a subject matter for a wide range of branches of mathematics (hopefully, all of them). The purpose of designating a particular proof system as a foundation is that it is then used as a setting in which one researches all these various branches, regarding proofs in the system about the relevant structures not just as derivations – empty of content – but as establishing their conclusions as true, as genuine proofs in the sense discussed in section VI.2. When Fermat's last theorem, the Poincaré conjecture or the Green-Tao theorem are proved in set theory, for instance, one does not just say that they have been proved in set theory but that they are theorems, that they have been proved, *tout court*: our current choice of set theory as a foundation means that results proved in set theory are taken as established, as truths.

Thus when deciding on a foundation for mathematics we are seeking a proof system which can provide a home for as many branches of mathematics as possible, by allowing us to (genuinely) prove statements about the structures which those branches

study. There is nothing in principle I think to bar us from studying different branches of mathematics in different proof systems, though one might then have to be careful when discussing a result to make clear which proof system it was proved in: for instance if proof system T_1 allowed the construction of certain kinds of topological spaces which proof system T_2 did not, then one would have to be careful when reasoning using the methods of T_2 not to appeal to any results proved in T_1 about the existence and properties of those kinds of topological spaces. In this kind of case, the univocal notion of proof in mathematics would be lost – it is the choice of a single foundational proof system for mathematics that makes this univocal notion of proof possible.[18] Apart from the condition that it provide a home for as many branches of mathematics as possible, I believe the key property we want from a choice of foundation is just that it provide methods of proof that are maximally powerful and flexible – that deciding conjectures by these methods is as easy for mathematicians as it can be. In fact this second property is closely related to the first, as one of the main ways that a proof system can make results easier to establish is by allowing the construction of all manner of novel objects and structures, which can then allow new methods of attacking old problems (as well as providing the subject matter for new branches of mathematics).

This account of what we want from a foundation tallies closely with Maddy's comments on the subject. Indeed a potential foundation allowing structures studied by the various branches of mathematics to be defined and constructed, and results about them proved, are the key properties underlying Maddy's list of the benefits of a foundation (Maddy 2011, pp. 33–34).[19] She discusses that a foundation can "give explicit meaning to questions of existence and coherence", via the test of whether a certain kind of object/structure can be constructed in the proof system – a role which a proof system can only play if it has suitably rich methods for carrying out such constructions of novel

[18] At least in modern rigorous mathematics, where proof is always founded on some underlying formal proof system.

[19] Maddy is describing the role of set theory as a foundation, but her comments would apply to any suitable foundational system.

entities; "make previously unclear concepts and structures precise", which requires that the resources of the proof system for phrasing definitions be sufficiently expressive; and "facilitate interconnections between disparate branches of mathematics", as discussed above. Then the crucial property of using a foundation to "formulate and answer questions of provability and refutability" by "investigating what does or doesn't follow from the axioms of the theory" implicitly assumes that we are accepting the methods of the proof system as constitutive of what counts as a (genuine) proof, about the various structures we are interested in – otherwise these are purely formal properties of the proof system, of no especial wider interest. Finally, using a foundation to "identify perfectly general fundamental assumptions that play out in many different guises in different fields" is also only possible if we are accepting the methods of the proof system as valid for establishing results about the subject matter of the different branches of mathematics, as is "open[ing] the door to new strong hypotheses to settle old open questions".

In fact this account even appears to be compatible with Awodey's remarks about set theory, despite him being an avowed anti foundationalist. To clarify what he means by his if-thenist view, he describes one form of if-thenism that he intends to distance himself from (Awodey 2004, p. 10): that in which the laws, axioms and rules of the foundational system are taken as conditional, so that for instance if one proves in set theory a conditional "if A, then B", this result is reported as meaning that "if the laws, axioms, and rules of the system are true and correct, then if A, then B" (ibid., p. 10). He discusses that the argument against this form of if-thenism is that it makes all theorems of mathematics hypothetical: "We may never know whether the axioms of ZFC are true, or they may even be inconsistent, so it will not do to carry them along as conditions on every theorem" (ibid., p. 10). Instead, in his form of if-thenism, "the conditions are rather of the kind 'if G is a finitely generated abelian group', not 'if the axioms of ZFC are true' "(ibid., p. 10). On his version of if-thenism, the antecedents of the conditionals are not hypotheses, but rather "serve to specify the range of application of the subsequent

statement" (Awodey 2004, p. 10). Thus in terms of the vocabulary used in the discussion of interpretations of mathematics from section VI.3, Awodey's view appears to be more like a form of eliminative structuralism than if-thenism (though the views are obviously closely related) – which is consonant with his description of it as a form of structuralism. Regarding the reasoning used to derive results, he says that "establishing any 'if ... , then ... ' implication requires some tacitly assumed methods of reasoning, from simple chains of equations, to, say, ZFC" (ibid., p. 10): the principles of set theory may thus be assumed when establishing results about different kinds of structures, and are not taken to be hypotheses on which those results depend. Thus the role of set theory on his view is essentially its role as a foundation as described here – it is a body of reasoning that may be used without comment when establishing mathematical results about the kinds of structures we are interested in.

Now we combine these remarks on foundations with the discussion of genuine proof and interpretations from section VI.2 and section VI.5. For a proof system T to be a setting for genuinely proving results about structures of some kind S of structure of which we can obtain a faithful real world characterization, it needs to be the case that proof in T is factive for statements about such structures, so that if ϕ is a statement about such structures then $T \vdash \phi^T$ implies ϕ^I where I is a chosen interpretation of ϕ (with ϕ^T the translation of ϕ into the language of T). Call this condition the factivity of T with respect to I. Then for such an I to be acceptable, it needs to satisfy the eliminative constraint, so that whenever ϕ is a definable generalization about structures of this kind, we have that ϕ^I implies ϕ^{Elim}. Combining these, it follows that for T to be a setting for genuinely proving results about structures of kind S it needs to be the case that whenever ϕ is a definable generalization about S's, $T \vdash \phi^T$ implies ϕ^{Elim}.

This holding for generalizations of this form about some kind of structure (of which we can obtain a faithful real world characterization) is the condition of soundness for T with respect to that kind of structure, though we should determine more precisely what

kinds of generalizations are included, by specifying what logic and language we have in mind. For instance one could define T to be arithmetically sound when this holds of all first order (or second order) arithmetic statements, define T to be real-analytically sound when this hold for all statements in the language of the complete ordered fields in monadic second order logic, define T to be group theoretically sound when this holds for all statements in the language of the theory of groups in monadic second order logic, and so on.

Actually it appears plausible that this condition of soundness holding for a kind of structure (of which we can obtain a faithful real world characterization) is equivalent to the condition that there be an acceptable interpretation I of statements about that kind of structure for which proof in T is factive. Indeed if real examples of this kind of structure do exist, then the eliminative structuralist interpretation is non vacuous, and arguably is thus an acceptable interpretation (it trivially satisfies the eliminative constraint), and it is one with respect to which T is factive – this being essentially the definition of soundness. On the other hand if real examples of this kind of structure do not exist, then the eliminative constraint for this case is trivial, and presumably one could then regard the deductivist interpretation based on proof system T as acceptable, for which proof in T is trivially factive.

Now when deciding on a foundation for mathematics, we are seeking a proof system which will be a home for as many branches of mathematics as possible, by allowing us to define and construct the structures they study, and (genuinely) prove results about these structures. As just discussed, being able to genuinely prove statements about kinds of structures for which we can obtain faithful real world characterizations requires that the proof system be sound with respect to those kinds of structures – and soundness is also plausibly a sufficient condition for this. Thus when seeking a proof system to be a foundation for mathematics, we should be seeking one which is sound for as wide a variety of kinds of structures as possible.

In principle I think it is possible that one could settle on a (limited) foundation in the form of a proof system which allowed proofs about certain kinds of structures to be regarded as genuine proofs, but not about other kinds of structures. For instance, perhaps we could argue that a certain proof system T is sound for deriving first order statements about structures axiomatizable in first order logic, such as groups and graphs, and also sound for first order statements of arithmetic, but be unable to argue that it is sound for statements about the real numbers, of statements in plural logic about groups, graphs or the natural numbers. In this case we could potentially regard T as a kind of limited foundation, in which we can genuinely the appropriate kinds of statements about groups and graphs and the natural numbers, but regard other kinds of statements proved in the theory as having a purely instrumental role within the theory, rather than as being established as genuine facs. One would then have to be clear which kinds of results were which, though – some being genuinely established truths about a kind of structure, others being merely derived results which might have applications within the proof system, but not outside of it – so that in applications of mathematics only the genuinely established results were used, and applied to deduce facts about structures in reality or in idealized models.

Similarly one could conceivably pick on a kind of foundation which only allowed some limited class of statements about a given kind of structure to be regarded as being genuinely proved – for instance Π_1 statements of first order arithmetic. This would need to be a well defined, general class of statements though. One couldn't for instance take one's "foundation" to be the theory $\{\phi_{\mathrm{GC}}\}$ where ϕ_{GC} is the Goldbach conjecture, and use this to trivially "prove" the Goldbach conjecture, arguing that this is a sound theory since ϕ_{GC} is likely true as a fact about the natural numbers (Paseau 2015 discusses the wealth of evidence we have for the Goldbach conjecture, arguing that we are "virtually certain" of it, though lacking a proof). Also one can't try to say for instance that the arithmetic statements which are genuinely provable in T are the "natural" ones, since it

needs to be clear or easy to establish whether a statement is in the target class or not; and we have no criteria or decision procedure for what naturality amounts to here, so that it may well be arguable – in a way that is difficult or impossible to resolve – whether a statements belongs to this class or not. Thus we would not know which arguments we should actually be regarding as proofs – as it may not be clear whether their conclusions are "natural" – which is antithetical to the notion of proof (recall that as discussed in section I.5, a key feature of the modern rigorous notion of proof is that it provides an effective procedure for resolving disputes about the validity of proofs). Of course coming up with some criteria, for instance a syntactic one, for what is "natural" would resolve this issue, giving a well class of genuinely provable statements.

To conclude the section, it is worth noting that even if we forget about the discussion of interpretations from section VI.5, there is an attraction just in the direct idea that a foundation should be sound, for as many kinds of structures as possible, so that proving generalizations about those kinds of structures in the proof system correctly informs us of properties of real examples of them. This could be an attractive view of foundations for a fictionalist, giving an instrumental use for a foundation for mathematics (in a similar vein to the advocacy of the notion of conservativeness by Field 1980; 1989). However as discussed in section VI.3, since the deductivist and eliminative structuralist interpretations of mathematics are available, I think fictionalists need to give more sustained arguments for why the face value platonist interpretation of mathematics is correct if their view is to be an attractive one.

7 Arguing for soundness

We now briefly consider how one might argue for the soundness of a proof system. First we consider the relation between soundness and consistency, and then consider how extrinsic or inductive evidence might be used to argue for soundness.

As a technical note, we will assume in this section that we have the resources to talk about truth of mathematical statements as interpreted in terms of real structures.[20] For example, if we have a real infinite sequence A, with initial element a and successor function S_A, then we saw in section VI.4 that we can define addition and multiplication on such a sequence, and thus can interpret any statement ϕ of arithmetic as a statement ϕ^A about this sequence; and we will assume that we have the resources to talk about denotation of arithmetic terms in A, and truth and satisfaction of arithmetic statements, with $A \vDash \phi$ used to denote the satisfaction of arithmetic statement ϕ by the sequence A (the details of this how this might be set up will be skipped over). We note a few technical facts about this. Firstly, we obtain that each numeral – which may be an object of some theory of syntax, rather than an element of A – denotes an element of A under this interpretation; and we can argue by induction on elements x of A that if $x \in A$ then x is denoted by some numeral (using the open-endedness of induction along A). A useful fact we thus obtain is that if $\phi(x_1, \ldots x_n)$ is an open arithmetic formula then $A \vDash \forall x_1 \ldots x_k \phi$ iff for all numerals $n_1 \ldots n_k$, $A \vDash \phi(n_1|x_1, \ldots n_k|x_k)$, and $A \vDash \exists x_1 \ldots x_k \phi$ iff there are numerals $n_1 \ldots n_k$, $A \vDash \phi(n_1|x_1, \ldots n_k|x_k)$. Also, PA is sound for A – all axioms of PA hold when interpreted in PA, and arguments in first order logic preserve truth, so that if $\text{PA} \vdash \phi$ then $A \vDash \phi$.

Now, to the main arguments of the section. A first key point is that soundness is a stronger condition than consistency. Assuming that PA is consistent, we can obtain a consistent proof system T' by adding in the statement that PA is inconsistent, $\neg\text{Con(PA)}$; but we can argue that this proof system T' is arithmetically unsound. Indeed suppose we have a real simply infinite sequence A. If $A \vDash (\neg\text{Con(PA)})$, then we have

$$A \vDash (\exists x \exists y \ x \text{ is a formula and } y \text{ is a derivation in PA of } (x \wedge \neg x))$$

[20] This does not involve any self referential uses of the truth predicate so should be uncontroversial.

and thus (by the discussion in the first, technical paragraph) that there is some numeral n and some first order arithmetic statement ϕ such that

$$A \vDash (n \text{ codes a derivation in PA of } \phi \wedge \neg\phi)$$

and thus that $A \vDash (\text{PA} \vdash \phi \wedge \neg\phi)$, so that by soundness, $A \vDash \phi \wedge \neg\phi$, which is impossible. Thus the assumption that $A \vDash (\neg\text{Con(PA)})$ must be false, so that the proof system T' is arithmetically unsound.

The example of this theory T' may be felt to be unnatural, so further examples may be helpful. As seen in Appendix D, we can define a number of concepts from standard real analysis in terms of a complete ordered field (as axiomatized in section VI.4), and can state what it is for a plurality of elements of a complete ordered field to be analytic, or $\mathbf{\Sigma}^1_2$, or to have the Baire property, or the perfect set property, or to be Lebesgue measurable; and we can also interpret talk of games on $\mathbb{N}$ of length $\mathbb{N}$, and what it is for such a game to be determined (for one player or the other to have a winning strategy). We can thus ask whether every $\mathbf{\Sigma}^1_2$ plurality of reals has the Baire property, or the perfect set property, or is Lebesgue measurable. To answer this question, we actually need to supplement ZFC with additional axioms. As shown by Solovay, if we add an axiom stating the existence of a measurable cardinal to ZFC then we can prove that all $\mathbf{\Sigma}^1_2$ subsets of $\mathbb{R}$ do have the Baire property, the perfect set property, and are Lebesgue measurable (see Kanamori 1997 §14, specifically 14.3 and 14.10 – p.179 and p.184). Then if we write MC for the statement that there is a measurable cardinal, then if ZFC + MC is sound for the reals then we can deduce that for any real example of a complete ordered field, all $\mathbf{\Sigma}^1_2$ pluralities have the Baire property, the perfect set property, and are Lebesgue measurable. On the other hand if we add the axiom $V = L$ to ZFC, we get the opposite answers: One can prove that there is a Δ^1_2 set which is not Lebesgue measurable and does not have the Baire property, and there is a Π^1_1 set which

does not have the perfect set property (see Kanamori 1997, 13.10 and 13.12, pp.169–170). It follows that there are $\mathbf{\Sigma}^1_2$ sets with all these properties. Thus if ZFC + $(V = L)$ is sound for the reals, we can deduce that for any real example of a complete ordered field, there is a $\mathbf{\Sigma}^1_2$ plurality which does not have the Baire property, one which does not have the perfect set property, and one is not Lebesgue measurable. Thus if both ZFC + MC and ZFC + $(V = L)$ were sound for the reals, we would obtain that for any real example of a complete ordered field, it is the case that every $\mathbf{\Sigma}^1_2$ plurality has the Baire property, but also the case that there is a $\mathbf{\Sigma}^1_2$ plurality which does not have the Baire property (and similarly for the perfect set property and Lebesgue measurability); thus we would obtain that there cannot in fact any real examples of complete ordered fields, so that the eliminative constraint and the condition of soundness for the reals are vacuous. This is despite ZFC + MC and ZFC + $(V = L)$ both being consistent, as far as we know. Thus we have two (apparently) consistent theories which cannot both be sound for the reals, unless soundness for the reals is a vacuous property.

The continuum hypothesis is also interpretable in terms of any real example of a complete ordered field – it just states that any plurality of elements either injects into $\mathbb{N}$ (i.e. the copy of $\mathbb{N}$ in the field) or is bijective with the whole field. Thus the theories ZFC + CH and ZFC + $\neg$CH, which are relatively consistent with ZFC, cannot both be sound for the reals, unless again soundness for the reals is a vacuous property. For a final example, Appendix D shows how one can interpret talk of subpluralities of $\mathbb{N}^{\mathbb{N}}$ in terms of any real example of a complete ordered field, and can regard such subpluralities as the target set of a game on $\mathbb{N}$ of length $\mathbb{N}$, where each of two players takes it in turn to select natural numbers, and the first player wins iff the resulting sequence is in the target set. One can define what it is for such a game to be determined – for one player or the other to have a winning strategy – and can thus phrase the statement that all such games are determined, in terms of any real example of a complete ordered field. Then if AD is the axiom of determinacy, the soundness of the apparently consistent theory ZF + AD

for the reals would imply that for any real example of a complete ordered field, every such game is determined, whereas the soundness of ZFC for the reals would imply that for any real example of a complete ordered field, there is a game which is undetermined (this being a standard consequence of the axiom of choice). Thus again, ZF + AD and ZFC cannot both be sound for the reals unless soundness for the reals is vacuous.

From these examples, we can already see that there is no way to argue for the soundness of a general proof system by a quick assessment of its superficial plausibility, or coherence. The proof systems ZFC + CH and ZFC + ¬CH are both plausible, and both appear to be coherent, but cannot both be sound (unless no complete ordered fields exist). The same goes for ZFC + MC and ZFC + $(V = L)$, and for ZFC and ZF + AD (which all have at least some plausibility). When establishing the soundness of a proof system, some sort of deeper reason than just its apparent plausibility and coherence will have to be appealed to. Incidentally these kinds of examples tell against Shapiro's claim that any coherent theory characterizes a structure, or (non empty) class of structures, which he describes as "[t]he main principle behind structuralism" (Shapiro 1997, p. 95). The second order versions of all these theories are perfectly coherent, so by Shapiro's principle it follows that there is a structure satisfying the $\mathrm{ZFC}^2 + \mathrm{CH}$, and one satisfying $\mathrm{ZFC}^2 + \neg\mathrm{CH}$, for instance. But in the former there is a substructure $\mathbb{R}_{\mathrm{CH}}$ which is a model of the second order theory of complete ordered fields which satisfies the continuum hypothesis, and in the latter there is a substructure $\mathbb{R}_{\neg\mathrm{CH}}$ which is a model of the second order theory of complete ordered fields which does not satisfy the continuum hypothesis. Moreover by the categoricity of the second order theory of complete ordered fields, we have that $\mathbb{R}_{\mathrm{CH}}$ and $\mathbb{R}_{\neg\mathrm{CH}}$ are isomorphic; thus we obtain a contradiction. Thus either Shapiro's principle – that all coherent theories characterize a structure, or (non empty) class of structures – is false, or apparently coherent theories may fail to be coherent. But in the latter case, coherence cannot do the work Shapiro needs it to do, as we then cannot know that ZFC^2 for instance, or the second order theories of complete ordered

fields or of arithmetic, are coherent, any more than we can know that $\mathrm{ZFC}^2 + \mathrm{CH}$ and $\mathrm{ZFC}^2 + \neg\mathrm{CH}$ (or $\mathrm{ZFC}^2 + \mathrm{MC}$, $\mathrm{ZFC}^2 + (V = L)$, ZFC^2 and $\mathrm{ZF}^2 + \mathrm{AD}$) are coherent.

Returning to the main topic of the section, we can argue that consistency does get us a certain amount of soundness though – soundness for various finitary structures and statements. For instance if there really exist objects $a_1, \ldots a_{200}$ with $R_\times$ giving a group structure on them, a well defined, precise relation, then in ZFC will be able to define a set $X = \{x_1, \ldots x_{200}\}$ with ternary relation $\times_X$ such that for each i, j, k, $\mathrm{ZFC} \vdash x_i \times x_j = x_k$ iff $R_\times(a_i, a_j, a_k)$, and $\mathrm{ZFC} \vdash x_i \times x_j \neq x_k$ iff $\neg R_\times(a_i, a_j, a_k)$, and thus such that $(X, \times_X)$ is a group. Then if ZFC is consistent, and ZFC proves that all finite groups satisfy group theoretic property $\mathcal{Q}$, then is must be the case that $(X, \times_X)$ satisfies property $\mathcal{Q}$ (ZFC would otherwise be inconsistent), and thus that the $a_1, \ldots a_{200}$ under operation $R_\times$ satisfy $\mathcal{Q}$; a metatheoretic induction shows that if $\phi(y_1, \ldots y_n)$ is an open formula of group theory, then $\phi(a_{i_1}, \ldots a_{i_n})$ holds iff $\mathrm{ZFC} \vdash \phi(x_{i_1}, \ldots x_{i_n})$, and any single instance of this equivalence follows without requiring metatheoretic reasoning. One gets soundness for finite finite graphs, finite fields, finite topological spaces and so on in the same way.

We can also argue that consistency delivers soundness for Π^0_1 statements of arithmetic – statements of the form $\forall x_1 \ldots \forall x_k \phi$ where ϕ contains only bounded quantifiers. Suppose that we have a real simply infinite sequence A, as discussed initially. We have seen that if $\mathrm{PA} \vDash \phi$ then $A \vDash \phi$. Also, given any closed formula ϕ in the language of arithmetic containing only bounded quantifiers, it is easy to see that $\mathrm{PA} \vdash \phi$ or $\mathrm{PA} \vdash \neg\phi$. Thus suppose that $\mathrm{ZFC} \vdash \forall x_1 \ldots \forall x_k \phi$ where ϕ has only bounded quantifiers, and that $A \vDash \neg\forall x_1 \ldots \forall x_k \phi$. Then for some numerals $n_1, \ldots n_k$, $A \vDash \neg\phi(n_1|x_1, \ldots n_k|x_k)$. But then we also have that either $\mathrm{PA} \vdash \phi(n_1|x_1, \ldots n_k|x_k)$ or $\mathrm{PA} \vdash \neg\phi(n_1|x_1, \ldots n_k|x_k)$, but the former would imply $A \vdash \phi(n_1|x_1, \ldots n_k|x_k)$, which would give a contradiction. Thus we must have $\mathrm{PA} \vdash \neg\phi(n_1|x_1, \ldots n_k|x_k)$, and thus $\mathrm{ZFC} \vdash \neg\phi(n_1|x_1, \ldots n_k|x_k)$, so that ZFC is inconsistent. Thus if ZFC is consistent, and $\mathrm{ZFC} \vdash \forall x_1 \ldots \forall x_k \phi$ where ϕ has

only bounded quantifiers, then $A \vDash \forall x_1 \ldots \forall x_k \phi$, as claimed. These soundness properties ZFC has if consistent will also hold of other general proof systems if consistent, such as ETCS, homotopy type theory, and PA^2.

A stronger property than consistency is Σ^0_1 soundness, the property of soundness for Σ^0_1 statement of arithmetic; Σ^0_1 statements being those of the form $\exists x_1 \ldots \exists x_k \phi$ where ϕ has only bounded quantifiers. We can argue that if ZFC (as an example) is Σ_1 sound, and any infinite sequence A as above exists, then ZFC is sound for first order statements about groups. Indeed suppose that ϕ states that all groups have property $\mathcal{Q}$, where $\mathcal{Q}$ is a first order definable properties of groups, and that $\mathrm{ZFC} \vdash \phi$. Then if we let T_G be the theory of groups as defined via some arithmetic coding in ZFC, and ϕ^{T_G} be the statement of property $\mathcal{Q}$ in the language of T_G, then it follows that $\mathrm{ZFC} \vdash (T_G \vDash \phi^{T_G})$, and thus since ZFC proves the completeness of first order logic that $\mathrm{ZFC} \vdash (T_G \vdash \phi^{T_G})$. But this is a Σ^0_1 statement of arithmetic, and so we obtain that $A \vDash T_G \vdash \phi^{T_G}$, and thus by the soundness of first order logic (applied to first order logic as defined using arithmetic coding in terms of sequence A), that if R_G is any real example of a group structure then R_G satisfies property $\mathcal{Q}$. One similarly obtains that Σ^1_0 soundness delivers soundness for first order statements about graphs, or rings, or fields, or partial orders, provided one assumes that there is an infinite sequence A as here, or makes some other assumption that allows this argument to go through.

There is a reasonable case that by exploring the consequences of a proof system, we can reassure ourselves that there are no apparent inconsistencies, and become highly confident that further exploration will not stumble across any. Potter (2004, p. 35) doubts this point has been reached for set theory, as the theory has not been pushed "to its limits", as do Linnebo and Pettigrew (2011, p. 248), as contradictions have not been sought in the parts of the theory "closest to paradox". Nonetheless neither deny that in principle one could explore a theory thoroughly enough that the chance of encountering a future inconsistency was established to be very low. If this could be done, then one

could argue from the above that one could be confident not to encounter an unsoundness concerning finite groups, finite graphs, finite fields and so on, or concerning Π^0_1 statements of arithmetic.

The prospects of using extrinsic or inductive evidence to argue for any more soundness than this appear to be very weak, however. Suppose for instance that one tried to argue for Σ^0_1 soundness by inductive evidence, arguing that whenever some proof system T proves a Σ^0_1 statement, this statement actually holds of any infinite sequence that exists. We currently cannot confirm the existence of any physical infinite sequences, so have no way of physically checking whether any Σ^0_1 statement holds of them. The only way to "confirm" a Σ^0_1 statement of arithmetic $\exists x_1 \dots \exists x_k \phi$ would be to find, either by hand or by computer, numerals $n_1 \dots n_k$ such that $\phi(n_1|x_1, \dots n_k|x_k)$ holds. Here $\phi(n_1|x_1, \dots n_k|x_k)$ is a decidable statement, which can be checked in a routine (though perhaps time consuming) way using basic arithmetic properties. $\phi(n_1|x_1, \dots n_k|x_k)$ being checkable in this sense is equivalent to it being the case that $\text{PA} \vdash \phi(n_1|x_1, \dots n_k|x_k)$; and we call a choice of numerals $n_1, \dots n_k$ such that $\text{PA} \vdash \phi(n_1|x_1, \dots n_k|x_k)$ a *witness for* $\exists x_1 \dots \exists x_k \phi$ *in PA*. If such a witness does exist, then $\exists x_1 \dots \exists x_k \phi$ does hold in any infinite sequence A, since then we have $A \vDash \phi(n_1|x_1, \dots n_k|x_k)$, and thus $A \vDash \exists x_1 \dots \exists x_k \phi$. One could potentially try to confirm the Σ^0_1 soundness of a proof system T by checking for witnesses in PA, seeing if when $T \vdash (\exists x_1 \dots \exists x_k \phi)^T$ there are numerals $n_1 \dots n_k$ such that $\text{PA} \vdash \phi(n_1|x_1, \dots n_k|x_k)$. The problem is that this is not actually confirming the statement that T is Σ^0_1 sound, but is instead evidence for the claim that if T proves a Σ^0_1 statement, then there are witnesses for this statement in PA; and if T extends PA then this claim is false, since there are examples of such ϕ for which $\text{PA} \vdash \exists x_1 \dots \exists x_k \phi$ – and thus $T \vdash (\exists x_1 \dots \exists x_k \phi)^T$ – but for all numerals $n_1, \dots n_k$, $\text{PA} \nvdash \phi(n_1|x_1, \dots n_k|x_k)$ (assuming PA is consistent).[21] There is no way that this kind of evidence can give

[21] The following two examples were given by Joel David Hamkins on Mathoverflow (https://mathoverflow.net/questions/190711/existential-statement-without-witness). Firstly, PA proves that there is a number n such that if there is no proof of a contradiction in PA of length at

us reason to believe that if T proves a Σ^0_1 statement for which there are no witnesses in PA, such a statement is actually true of simply infinite sequences that exist – and so since there will be provable Σ^0_1 statements without a witness in PA for any general proof system T, this kind of evidence cannot deliver the Σ^0_1 soundness of a general proof system.

The prospects of confirming other instances of soundness by inductive evidence are similarly dim (or dimmer). For most kinds of mathematical statements, we have no hope of checking whether they hold by examining physical examples of the relevant structures. Consider the statement that every $\mathbf{\Sigma}^1_2$ subset of a complete ordered field has the perfect set property. As discussed in section VI.4, in our present state of knowledge we cannot confirm that the physical universe is infinite in any sense, let alone whether it contains any complete ordered fields. We discern properties of spacetime by observing how particles behave, and cannot even determine the position of a particle to an arbitrary degree of accuracy – so even experimentally determining whether a line segment in space really satisfies the completeness axiom, or only completeness for some restricted class of subsets (perhaps the open or closed subsets) is not a problem we have any prospect of being able to resolve. When physically indicating a subregion of such a line segment, we are again limited to a finite degree of accuracy, so it seems implausible that we could even reliably refer to a particular $\mathbf{\Sigma}^1_2$ subset of it, much less determine whether is satisfies the perfect set property. The same goes for other mathematical statements. Consider for instance the statement that every subgroup of a finitely generated abelian group is finitely generated. We are unable to even pick out any infinite physical groups, not knowing whether the universe is infinite or not; and even if we could, trying to examine their subgroups by physical experiment would be a thankless task.

most n, then there is no proof of a contradiction in PA at all; since if PA is consistent, then the antecedent is true, and if PA is inconsistent, then one can take n to be the length of shortest proof of a contradiction in PA. But PA does not prove that this property holds of any particular number (unless it is inconsistent), since then it would prove its own consistency. For a second example, if σ is any statement unprovable in PA, and we let $\phi(x)$ be the statement $(x = 0 \wedge \sigma) \vee (x = 1 \wedge \neg\sigma)$, then PA proves $\phi(0) \vee \phi(1)$, and thus proves $\exists x \phi(x)$, but does not prove $\phi(n)$ for any numeral n.

One could perhaps try to confirm such mathematical statements by finding established physical theories in which they play an essential role. As far as I'm aware, statements like the first – that every $\mathbf{\Sigma}^1_2$ subset of a complete ordered field has the perfect set property – currently play no physical role, so this option will not help with them. Many mathematical generalizations are used in physical theories however, and it is possible that the second statement – that every subgroup of a finitely generated abelian group is finitely generated – or ones like it play some role in certain physical theories (I do not know). All this could establish though is that the theorem does apply (for instance) to all groups of which it is assumed to hold by the theory – not that it actually holds of all physical examples of groups which exist (let alone all groups that exist). Anyway, when a mathematical theorem like this is used in physics, it is generally taken for granted that it is true; physicists will rarely set up competing theories, one in which a mathematical theorem is assumed to be true, one in which it is assumed to be false, and the two theories tested against each other. Since mathematical theorems are not tested as empirical claims in the same way as other aspects of the theory, instead playing a regulative role in setting it up, it is hard to see how they could be confirmed in the same way as those more empirical aspects. Additionally, so much of physics itself is still so uncertain – what for instance should we infer from quantum mechanics about the nature of the world? what is the real nature of spacetime? – and lacking in confirmation that it is hard to see how the more obscure, mathematical parts of theories could be definitely confirmed in this way. Finally, many mathematical theorems are simply not used in physics in any way currently, so the prospects of establishing soundness for all statements about any kind of structure in this way are very dim.

Another option is, as with the case of Σ^0_1 soundness, to try to settle these kinds of mathematical questions by resorting to evidence from within mathematics itself. One general method for trying to do this starts with a base proof system T which we believe, and tries to confirm a proof system T' which extends T by checking whether when we

can prove a statement in T', we can actually prove it already in T. Here for instance T might be the theory of groups in first order logic, sound for reasoning about groups, or the theory PA^2, sound for reasoning about simply infinite sequences. In cases like the former, this may be a valid strategy: perhaps one could indeed argue for the soundness of a general proof system T' by seeing whether when it proves first order statements of group theory, these are actually provable in the first order theory of groups. However in general, this style of reasoning runs into the same problem as in the case of Σ^0_1 arithmetic soundness: what we're really confirming by this kind of argument is not the claim that T' is sound for reasoning about the relevant kind of structure, but the claim that if $T' \vdash \phi^{T'}$ then $T \vdash \phi^T$, and this is false for many classes of structures and many general proof systems T'. This kind of argument tells us nothing about the soundness of T' for statements about the relevant kind of structure that are not provable in T, which often will exist.

What we need is some way to use the base proof system T to confirm statements not provable in T. One way to attempt this is to try to confirm statements of the form $\forall x_1 \ldots x_n \phi$ by verifying that $T \vdash \phi(t_1|x_1, \ldots t_n|x_n)$ for as many terms $t_1, \ldots t_n$ as possible, or equivalently trying to disconfirm statements of the form $\exists x_1 \ldots x_n \phi$ by being unable to find terms $t_1, \ldots t_n$ such that $T \vdash \phi(t_1|x_1, \ldots t_n|x_n)$.[22] Then one could try to confirm a proof system T' extending T by seeing if consequences of T can be confirmed in this manner – where again T could be a proof system known to be sound with regard to some kind of structure, and T' a more general proof system. This is essentially the kind of evidence seen in a famous case study involving determinacy hypotheses described by D. A. Martin (1998), which is taken to be a paradigmatic example of how one can confirm theories by extrinsic evidence and investigation of their consequences – discussed for instance by Maddy (1997, p. 72; 2011, pp. 127,130). The problem with this kind

[22]Formally our language may not actually contain many terms, or ways of forming terms, so instead of $\phi(t_1|x_1, \ldots t_n|x_n)$ one could consider statements of the form $\forall x_1 \ldots x_n(\psi_1(x_1) \wedge \ldots \wedge \psi_n(x_n)) \Rightarrow \psi(x)$ where $T \vdash \exists! x_i \psi_i(x_i)$ for each i

of "confirmation" is that if we have an existential statement $\exists x_1 \dots x_n \phi$ independent of T, this kind of "evidence" will only serve to disconfirm it, and if we have a universal statement $\forall x_1 \dots x_n \phi$ independent of T, then this kind of "evidence" will only serve to confirm it. Indeed if a statement $\exists x_1 \dots x_n \phi$ is independent of T, then there will be no terms $t_1, \dots t_n$ for which T can prove $\phi(t_1|x_1, \dots t_n|x_n)$ – for any such terms, either T will prove $\neg\phi(t_1|x_1, \dots t_n|x_n)$, or it will leave this substitution instance undecided. Either way, whichever terms $t_1, \dots t_n$ we consider, we will find no evidence for the statement $\exists x_1 \dots x_n \phi$. This statement will automatically just accumulate disconfirming evidence on this approach. By contrast if a statement $\forall x_1 \dots x_n \phi$ is independent of T, then no matter which terms $t_1, \dots t_n$ we consider, it will always be the case that either T proves $\phi(t_1|x_1, \dots t_n|x_n)$, or T leaves the substitution instance undecided. Either way we find nothing to tell against the assumption that $\forall x_1 \dots x_n \phi$ holds, and may well discover evidence for it – so that $\forall x_1 \dots x_n \phi$ will only every accumulate confirming evidence on this approach. Thus we can choose to confirm or disconfirm a statement depending on whether we phrase it as a universal or existential statement. For instance we can confirm the statement "no set is intermediate in cardinality between $\mathbb{N}$ and $\mathbb{R}$" by seeking one, and being unable to find it – and thus confirm the continuum hypothesis; or we can confirm the statement "there is no bijection between ω_1 and $\mathbb{R}$" by seeking one, and being unable to find it, and thus confirm the negation of the continuum hypothesis. A sign of the weakness of this kind of evidence is that in practice, it is not typically taken to be convincing. If it was, then if we accept the axioms of ZF as governing the set theoretic hierarchy, we would have an overwhelming weight of evidence against the axiom of choice: there are many cases where we can prove $\exists x \phi(x)$ with the axiom of choice, but cannot show the existence of any example without it, for instance the existence of a right inverse to any surjection, the existence of a non measurable subset of the real line, the existence of a maximal ideal containing any proper ideal in a ring, the existence of a basis of the reals as a vector space over the rationals, and so on. We would indeed also

have a strong case against any existential statement independent of ZFC, such as the existence of an inaccessible cardinal – however hard we try in ZFC, we cannot prove the existence of such a cardinal, so by this line of reasoning we should reject the statement of its existence. Of course this is not a convincing argument, because of the weakness of this form of evidence.

There are a variety of other kinds of extrinsic evidence that have been appealed to by set theorists as evidence for a theory, as discussed by Maddy (1988a,b; 2011, §II.2, §V.3). None of the other forms appear to have much potential as a way of arguing directly for the soundness of a general proof system, though one could perhaps try to use extrinsic evidence of these kinds to argue for the truth of an axiom system for set theory for describing some domain of sets, and thus for the soundness of the axiom system.

This kind of possibility – of arguing for the soundness of a proof system via its truth with respect to some subject matter – is what we consider now. It appears to be by far the most plausible route to arguing for the general soundness of a proof system.

For instance, suppose one believes in the iterative conception of set. On this perspective sets are formed in a series of cumulative stages. Briefly, at the initial stage, all sets of individuals (things that aren't sets) are formed; then, all sets of things at that stage, or before (individuals or sets of individuals) are formed; then, all sets of things formed at *that* stage, or before, are formed; and so on. After the first infinity of stages, there is a stage consisting of all sets formed at any one of those stages; then there is a stage after that, at which all sets of things at that stage are formed; and so on (for more careful descriptions of this conception, see Boolos 1971 or Potter 2004, Chapter 3). One can argue that the sets formed in this manner collectively satisfy the various axioms of ZFC. Indeed Boolos (1971, pp. 224–231) shows this for the axioms of Zermelo set theory, hesitating only about extensionality; this because extensionality appears to be analytic of our concept of set, and Boolos has Quinean doubts about the notion of analyticity. However we can perfectly well take extensionality to be partly constitutive

of what we mean by "set", as Burgess (2004, p. 199) does, so that a cumulative hierarchy of objects not satisfying extensionality would not be a cumulative hierarchy of sets in our sense (and stating this does not require a definition of analyticity, or a sharp distinction between the analytic and the synthetic). Boolos believes that the iterative conception is neutral about the truth of the axiom of choice, but Paseau (2007, p. 34–35) argues that it flows naturally from a combinatorial understanding of the set formation process – and otherwise one can obtain it from a form of choice phrased as a logical principle in plural logic, or via a version of Hilbert's ϵ-operator for pluralities (Burgess 2004, §8). That leaves the axiom scheme of replacement. Boolos (1971, pp. 228–229) notes that this follows from a bounding or cofinality principle for stages, which he considers attractive, though he believes it to be a further thought than that contained in the basic form of the iterative conception. Otherwise Paseau (2007, p. 33) notes that the instances of replacement follow from a reflection scheme, which is justified by the idea that the set theoretic hierarchy is absolutely infinite and "transcends unique mathematical specification" – though Potter (2004, Chapter 13) view this as a separately motivated principle.

Suppose now that we believe in the iterative conception of set, and in fact are convinced by it, and convinced that all the axioms of ZFC hold of the resulting hierarchy. Actually, the theory we want is not a pure set theory, since the initial stage of the hierarchy consists of all individuals – all things that aren't sets – and our theory needs to be able to talk about these as well as about sets. Since we are regarding set theory here as being part of how we describe what the real world is like, we use the same background logic as discussed in section VI.4, using a combination of restricted predicative second order logic together with plural logic and double ancestral logic (one could also use the plural ancestral and plural double ancestral, though we won't need them). Then we modify ZFC into a theory ZFCU in this logic in which urelements are admitted. ZFCU can be formalized by introducing predicates Ure and Set into the language, with every

object satisfying exactly one of these, and where if $\mathrm{Ure}(y)$ then for all x, $x \notin y$. We can include any vocabulary which holds of individuals (such as "physical" or "spatially located"), provided perhaps that it is suitably sharply defined and determinate, and can adjust any theory T which holds of individuals (non sets) to this context by relativizing its quantifiers to the predicate Ure. We introduce into ZFCU an axiom stating that there is a set containing all urelements (formed at the first stage of set formation above). We modify extensionality to only hold of sets – urelements may be unequal even if they have the same members (i.e. both having no members).

An important point is that we extend the separation scheme in an open ended manner, so that for any formula ϕ formed from all the vocabulary available in this context – including second order variables, instances of the double ancestral, and any empirical vocabulary we have available – we admit the statement that

$$\forall x\, \exists y\, \forall z\, (z \in y \leftrightarrow ((z \in x) \wedge \phi)),$$

or its universal closure if ϕ contains parameters (we require in this scheme that y is not free in ϕ).[23] Extending separation in this manner is fully justified by the iterative conception of set, in exactly the same manner as the original separation scheme of ZFC is justified. Indeed as Boolos discusses, we obtain from the iterative conception of set an axiom scheme which he calls specification, whose instances are for each open formula $\psi(z)$ that for any stage s, there is a set consisting of just those objects[24] to which $\psi(z)$ applies that were formed before s (Boolos 1971, p. 223). In this scheme the vocabulary used to define ψ is irrelevant (as long, perhaps, as ψ is suitably sharply defined and determinate) – all that is needed is that it is something that objects can satisfy or fail to satisfy. Thus the specification scheme should be taken as open ended,

[23] This scheme does not have to be restricted to sets x, since if x is a urelement then we can take y to be the empty set, or another urelement.

[24] Boolos restricts his scheme to just sets here, rather than allowing individuals, but there seems to be no basis for this.

and extended to include all vocabulary that we have available, and Boolos's argument for the separation scheme then extends immediately to an argument for the open ended version of separation (Boolos 1971, p. 226).

If one takes the cofinal or bounding approach to replacement that Boolos (ibid., p. 228) discusses, then the replacement scheme should also apparently be taken as open ended. If one argues like Paseau (2007, p. 33) for replacement via a reflection principle, then I think there is also a case for taking an open ended form of the reflection principle, leading to open ended replacement, though problems can arise from extending reflection principles in a naive manner and I am not certain of this route. Though there may be a case either way for an open ended form of replacement, we will not need it, and we will just take the replacement scheme in ZFCU to consist of its normal instances, not involving second order variables, plural logic, or the double ancestral (I do not think there is any objection to allowing the predicates Ure and Set in the scheme).

The other axioms of ZFC – power set, union, pair set, foundation and so on – are unchanged. Foundation takes the form that any set has an ϵ-minimal element, which may just be a urelement. In this context we call a set *pure* if its transitive closure only contains sets, rather than urelements, and call it *impure* otherwise.

Suppose then that we are convinced by this theory ZFCU as a theory of how (part of) the world is. Then we can argue straightforwardly for the soundness – with a caveat, to be discussed – of ZFC with respect to a wide range of structures. To see this, let T be a theory in monadic 100^{th} order logic in language L, and let ϕ be a statement L such that $\text{ZFC} \vdash (T \vDash \phi)$. We will argue that $\text{ZFCU} \vdash (T \vDash \phi)$ (we call this the transfer result). Indeed working within ZFCU we can define what it is for a set to be pure, as above, and prove for each axiom of ZFC that that axiom holds when relativized to the pure sets. Thus if $\text{ZFC} \vdash (T \vDash \phi)$ then we can prove in ZFCU that amongst the pure sets $T \vDash \phi$, i.e. that if $\mathcal{B}$ is an L-structure which is a pure set, then if $\mathcal{B}$ satisfies every axiom of T then $\mathcal{B}$ satisfies ϕ. Now let $\mathcal{C}$ be any L-structure satisfying the axioms of T,

possibly an impure set, with C its base set. In this context the proof that all sets have a cardinality – i.e. that there is a bijection between any set and some cardinal – goes through as usual, including for impure sets. In particular we can prove the existence of a bijection between C and some cardinal D. The latter is of course a pure set. Then we can transfer across the L-structure from $\mathcal{C}$ to D, defining an L-structure $\mathcal{D}$ on D which is isomorphic to $\mathcal{C}$, and moreover where $\mathcal{D}$ is a pure set. Thus since $\mathcal{D}$ satisfies every axiom of T (as $\mathcal{C}$ does), $\mathcal{D}$ satisfies ϕ, and so $\mathcal{C}$ satisfies ϕ, as required.

This gives us soundness for all the kinds of structures discussed in section VI.4, at least for set sized structures. Indeed suppose for instance that ϕ states that some definable property $\mathcal{Q}$ holds of all groups – where definable means definable in the combination of logic used in section VI.4, so the combination of restricted predicative second order logic with plural ancestral and double ancestral logic. Let ϕ^{ZFC} and ϕ^{ZFCU} be the translations of ϕ into the language of ZFC and ZFCU respectively. If we let T be the theory of groups in ω-th order logic, we can state $\mathcal{Q}$ as a property $\mathcal{Q}^T$ in the language of T that holds of a generic group (with plenty of wiggle room – this is why 100^{th} order logic was used, though it is huge overkill), and ϕ^{ZFC} is equivalent in ZFC to the statement that $T \vDash \mathcal{Q}^T$. Thus ZFC proves that $T \vDash \mathcal{Q}^T$, and so by the above, ZFCU $\vdash (T \vDash \mathcal{Q}^T)$. But again, $T \vDash \mathcal{Q}^T$ is equivalent in ZFCU to ϕ^{ZFCU}, and so ZFCU $\vdash \phi^{\text{ZFCU}}$. But then if $R_\times$ is any ternary relation which defines a group, as defined in section VI.4, and such its domain is "small enough" in that there is some set containing every element related by $R_\times$, then using open ended separation we can let G be the set of objects related by $R_\times$, and (again using open-ended separation) can define a set theoretic group structure on G corresponding to the relation $R_\times$. Then we can deduce from ϕ^{ZFCU} that this set theoretic group structure satisfies property $\mathcal{Q}$, and thus, finally, that $R_\times$ satisfies property $\mathcal{Q}$. Thus any definable generalization provable about groups in ZFC does actually hold of all real examples of set sized group structures.

This argument is very general, and applies to all the kinds of structures considered

in section VI.4, and much more widely – indeed 100^{th} order logic is huge overkill for axiomatizing almost any kind of mathematical structure (at least those I'm familiar with), and for being able to define properties that are definable in the combination of logics used in section VI.4. The number 100 here is of course arbitrary, and anyway even the use of the theory T here is not essential; if some kind of structure were not axiomatizable in this kind of logic, the argument would still go through as long as one could transfer the structure on an impure set C to a pure set D, and obtain an isomorphism between them.

The caveat to this argument for soundness via is that it only gives soundness for set sized structures. As noted right at the end of section VI.4, there is a case to be made that this is all the soundness we need. Indeed if we take set theoretic language to be meaningful, and to describe a genuine part of the world, as we are here, then a structure being the size of a proper class is a reasonable mathematical ground for not requiring our standard results to hold of it: we can take mathematical results to be implicitly restricted in their domain of applicability to structures that are not "too large" in this way – after all, when we characterize mathematical structures in set theory we do require them typically to be sets, and thus do rule out structures that are "too large" in exactly this sense. If this is the only variety of soundness we can obtain, it is certainly one we can live with, establishing as it does soundness for all structures consisting entirely of individuals (non sets) – since the collection of all individuals forms a set.

Additionally, we can still get soundness for certain kinds of structures even without the assumption that they are set sized. For instance if ZFC proves some arithmetic statement ϕ, and R is any relation defining the successor function of a simply infinite sequence, then using the double ancestral we can define a primitive recursive isomorphism between the objects related by R and the finite Von Neumann ordinals.[25] Then since ZFC proves ϕ, ϕ holds of all simply infinite sequences which are pure sets, and in particular

[25] We use here the open-endedness of induction on both sides, with open-endedness of induction for the Von Neumann ordinals following from the open-endedness of our separation scheme.

holds of the finite Von Neumann ordinals; and then since we have an isomorphism between them and the simply infinite sequence defined by R, we obtain that ϕ also holds of the latter. One can obtain the same result for complete ordered fields, with the key step being the definition of an isomorphism between a real example of a complete ordered field and some complete ordered field structure which is a pure set – using again the double ancestral to define a primitive recursive isomorphism between the natural numbers of the two complete ordered field structures, and then extending the isomorphism to the rest of the elements.

Also, if one takes a potentialist view of set theory, a view in which any objects whatsoever "could" have formed a set, in some sense, then one can argue for soundness even for proper class sized structures. Indeed suppose for instance that in ZFC we can prove that some definable property $\mathcal{Q}$ holds of all groups. Then given any relation $R_\times$ defining a group structure, one can argue that the domain of objects related by $R_\times$ could have formed a set, and that if it did then one could construct a pure set defining a group structure which is isomorphic to $R_\times$ (as above in the proof of the transfer result), and the latter would satisfy $\mathcal{Q}$, and so $R_\times$ would satisfy $\mathcal{Q}$ as well. Then one argues that whether $\mathcal{Q}$ holds of $R_\times$ is independent of whether its objects form a set or not, and thus since they could have formed a set, $\mathcal{Q}$ does hold of $R_\times$. There are a number of different attempts in the literature to formalize a potentialist view of set theory using a form of modal logic, such as Studd (2013) and Linnebo (2013); I will not attempt to give a more formal version of this argument using either of these frameworks, though that would be an interesting project.

This kind of justification of soundness via truth extends to more powerful set theories, if we are convinced by additional axioms. For instance if we believe an axiom stating the existence of an inaccessible or measurable cardinal, the same argument shows that adding this axiom to ZFC results in a theory which is sound for the same range of structures. Though extrinsic evidence appears to be insufficient to directly justify soundness in

general, as discussed above, a different route would be to try to use extrinsic evidence to justify the truth of the axioms, and thus their soundness. For instance if one is convinced by the iterative conception, one could then try to use extrinsic evidence to confirm what further axioms are true of the set theoretic hierarchy, for instance by examining their consequences, as discussed by Maddy (1988a,b; 2011). Soundness could then follow. This approach has its own problems, however, as discussed in section VI.9 when evaluating Maddy's recent remarks on set theory.

Other justifications of set theory can also lead to soundness, along the same lines. For instance Burgess (2004) gives motivation and arguments for the axioms of ZFC based on the limitation of size conception, as ultimately captured by a reflection scheme. He uses plural logic as his setting, and his justification for the axiom of choice goes via a form of choice for pluralities Burgess (ibid., §8), as mentioned above. If one is convinced by this limitation of size conception, as a description of the sets that really exist, then can again be led to phrase the theory – as just another part of our description of the world – in the background logic of section VI.4, including predicates Ure for individuals, and Set for set. One can argue that there is a set of all individuals, if one allows formulae involving the predicate Ure in the reflection scheme (and there seems to be no objection to this). One is led in this way to the same version of ZFCU as before. In particular one again has a strong justification for an open ended form of the axiom scheme of separation, in which one can separate subsets using any formula of our expanded language: indeed it is clear that the comprehension scheme for pluralities should be phrased in an open ended way – when discussing the objects satisfying some property, we need not be limited in what vocabulary we use to define that property – and open ended separation then follows from Burgess's single axiom of separation (ibid., p. 203). Having come to accept ZFCU as a genuine description of the way the world is, we can argue for the soundness of ZFC for a very wide variety of structures in exactly the same way as before.

That soundness argument uses replacement in an essential way – to find a pure set

structure $\mathcal{D}$ isomorphic to a given impure set structure $\mathcal{C}$. Without replacement, it is entirely possible that the set theoretic hierarchy does not extend high enough for this to be possible: if the rank of the cumulative hierarchy is small, and the set of urelements is sufficiently large, then there need be no pure set the same size as the set of all urelements (or subsets of it). Thus if one only believes a weaker set theory than ZFC – for instance if one does not believe that the iterative conception makes a strong case for the axiom scheme of replacement or reflection – then this kind of argument will not be available. However an argument like this was only needed because of the mismatch between the mathematical theory ZFC, in which there are only pure sets, and the theory ZFCU which describes sets as a feature of the real world, in which there are also individuals and impure sets. If instead of ZFC we worked in a set theory in which individuals/urelements were allowed, then there would be no such mismatch. For instance one could work in a first order version of the theory ZFCU described above, or in the theory ZU that Potter directly argues for from the iterative conception of sets Potter (2004, p. 72). Though I do not think Potter emphasises this point, as far as I'm aware in mathematics (outside of axiomatic set theory) one never actually uses the assumption that every object is a set, rather than an individual or urelement – for instance if reasoning about a group or topological space, one never uses the assumption that its elements are themselves sets (or for instance that real numbers are sets, except when constructing them and proving their basic properties). Thus – to my knowledge – a set theory with urelements like these would be just as suitable as a foundation for mathematics as their pure cousins (at least for the purposes of proving mathematical results about structures of interest).

For such theories we obtain soundness for set sized structures essentially immediately. Indeed suppose for instance that we accept Potter's theory ZU as our theory of sets, and that it proves that all groups have some definable property $\mathcal{Q}$. Let $R_\times$ be ternary relation defining some group structure, whose domain of objects is "small enough", i.e is contained in some set. Then by open-ended separation applied to the set of individuals,

we can form the set of all objects related by $R_\times$, and (again by open-ended separation) define a set theoretic group structure G on these objects corresponding to the relation $R_\times$. Then since ZU proves that all set theoretic groups have some definable property $\mathcal{Q}$, and we believe this theory is true of the sets that exist, then by the soundness of first order logic it follows that our group structure G has property $\mathcal{Q}$, as required.

For these kinds of soundness arguments to be convincing, we do actually have to be convinced of the truth of the set theory in question for describing the domain of sets. If we are merely positing ZFCU as one possible theory of sets, then we are under no obligation to take arguments like these seriously: if we are not convinced that ZFCU correctly describes the sets that exist – or that there are any sets – then when we argue that any structure made up of individuals is isomorphic to some pure set structure, and thus that any fact proved in pure set theory holds of it, these are just empty words. If we could just posit any set theory we like, and derive its soundness, then it would follow that ZFC + CH, ZFC + ¬CH, ZFC + MC, ZFC + $(V = L)$, ZF + AD and ZFC were all sound for the reals, but this is impossible as discussed earlier (unless no complete ordered fields exist).

The arguments for soundness via truth here all use the open-endedness of separation in an essential way, so that set theoretic surrogates can be found of any realizations of mathematical structures. As far as I can see, there is no obvious way to make the same kind of argument for other general proof systems that lack a separation scheme, such as ETCS and homotopy type theory. Even if one believed these theories to be true, on some interpetation, it is not at all clear (to me at least) how one would link vocabulary describing real world structures with the vocabulary of these theories, to establish that mathematical results proved in the theory actually hold of real world examples of structures. However for ETCS at least, I believe one can argue for soundness if one can first argue for the soundness of a membership based set theory, as above. Suppose for instance that one has established the soundness of ZFC, or of ZU. One can

give a fairly straightforward interpretation of ETCS as the theory of the category of sets in either case, and one can prove any generalization about some kind of structure in ETCS, one may well be able to transfer it across to a theorem about that kind of structure in ZFC or ZU respectively, and thus (via soundness of the latter) deduce that it holds of all real examples of that kind of structure. Thus if one does believe in the iterative conception of set, or the limitation of size conception, the soundness of ETCS may follow from that of ZFC or ZU. I do not think there is any need for the soundness of a general proof system to be "directly" justifiable, whatever that might mean: if one can justify the soundness of ETCS via the soundness of ZFC, then (as long as this argument is valid) that establishes ETCS as a valid candidate foundation, according to the requirements discussed here.

That still leaves the soundness of homotopy type theory unaccounted for. I am not an expert in the theory, so there may be ways to argue for this that I am not aware of. One possible option though is to use the ability to model homotopy type theory in set theory to again prove a relative soundness result, so that the soundness of set theory can be used to argue for the soundness of homotopy type theory. If homotopy type theory has an ω-model (a model in which the finite Von Neumann ordinals play the role of the natural numbers in the theory) then one could use this to argue for arithmetic soundness, at least. I would assume there was a known ω-model, but am not familiar enough with the relevant literature to know for sure. What these last two examples show though is that once we recognize the importance of soundness for deciding on a foundation for mathematics, it is not relative consistency results between proof systems that should most concern us, but relative soundness results: results showing that if some proof system T_1 is sound for certain kinds of structures, then so is some other proof system T_2.

8 Summary

Before drawing some implications of these arguments, it may help to briefly review what has been said. In section VI.2 it was noted that the term "proof" is used in different ways in mathematics: we talk of "proof in T", for a proof system T, where there need be no connotations that the conclusion of such a proof is actually true; but we also talk about proving a conclusion ϕ with no mention of a proof system T, where this latter notion of proof does establish its conclusion as true (or is assumed to). In order to discuss whether ϕ is true or not, we need to have some sort of interpretation of what ϕ means, and section VI.3 discussed various interpretations of mathematical statements that have been put forward. Section VI.4 then argued that – despite certain comments of Maddy (2011, p. 92) – we can generally give characterizations of mathematical structures in their own right, independent of general mathematical proof systems. Moreover we can give characterizations that have real world content – that can be satisfied, or fail to be satisfied, by real world objects. Section VI.5 then argued that if ϕ is a definable generalization about a kind of structure of which we can give such a characterization, stating that all such structures satisfy some property $\mathcal{Q}$, then an interpretation of ϕ should only interpret ϕ as true if it is the case that all real examples of that kind of structure actually do have property $\mathcal{Q}$. This was termed the eliminative constraint on interpretations. After that the key claim of the chapter was made in section VI.6, where it was argued that when seeking a potential foundation for mathematics, we should be seeking a proof system that can serve as a setting for as many branches of mathematics as possible, by allowing us to genuinely prove statements about the structures those branches study – where if the proof system in question is selected as a foundation, proofs in it are subsequently regarded as genuine proofs, as establishing their conclusions as true. But by the eliminative constraint, it is only possible for proofs in T to be genuine proofs of a class of statements on an acceptable interpretation if the proof system is

sound for those statements, where soundness of T for ϕ requires that if $T \vdash \phi^T$ then ϕ^{Elim} holds. Thus for a proof system to be a candidate foundation, it is necessary that it be sound for a wide range of structures, as many as possible. Finally section VI.7 considered a variety of ways to try to argue for soundness, and concluded that the only plausible way to argue for soundness for a wide variety of structures is in fact to argue for the soundness of ZFC via the iterative conception of set or the limitation of size conception – showing that some such conception establishes the axioms of ZFC to be true. The soundness of ETCS and homotopy type theory may then be (at least partially) justifiable by a relative soundness result.

9 Implications

We will now draw some implications of these arguments for various positions in the philosophy of mathematics, including those mentioned in section VI.1.

Firstly, as discussed in section VI.6, it appears that though Awodey (2004) is an avowed anti foundationalist, his view of the role of set theory is very similar to the view of foundations put forward here: it is a body of reasoning that may be assumed when establishing structural truths, truths of the form "all groups have property $\mathcal{Q}$" for instance. The difference between the perspective here and Awodey's is just that we consider what property a body of reasoning must have for it to be suitable to play this role – and that property is the property of soundness.

Next, we note that since soundness is not implied by consistency, the arguments of deductivists like Tait (2005) and Muller (2004) – claiming that all we want from our proof system is that it be consistent – cannot be right (unless they wish to argue for instance that there are no simply infinite sequences). What is missing from their accounts is the requirement that our mathematical results be compatible with the real examples of mathematical structures that exist. Tait's view of mathematical axioms, in

which we are apparently free to choose new axioms as we see fit, though we may be led to decide on certain more natural or attractive ones Tait (2005, pp. 91, 96–98, 294–295), is simply too liberal. There are constraints on our choice of proof system that go beyond consistency – if we want our mathematical results to be compatible with reality. As discussed in section VI.7, the notion of coherence is also not enough to justify a proof system, as there are pairs of apparently coherent theories whose soundness (for the reals) is mutually incompatible.

One implication for mathematical practice is that when judging potential foundations for mathematics, since we are concerned with their soundness rather than just their consistency, it is relative soundness results rather than relative consistency results that are most relevant – with relative soundness results with respect to set theory perhaps being the best way to justify the soundness of ETCS and homotopy type theory. When comparing for instance ZFC with ZFC + $(V = L)$, it is relevant that the latter is relatively arithmetically sound (as it has an ω model), rather than merely being relatively consistent.

Maddy (2011) is right to identify the property of mathematical depth as an attractive feature of a proof system, but in line with the arguments of this chapter, the first priority needs to be to find proof systems that are sound – with mathematical depth then being one kind of property that can be used to decide between sound systems. She is misled in her discussion of the role of a foundation by her assumption that mathematical structures automatically have the properties that advanced methods (i.e. set theory, and extensions of it) discern in them (ibid., p. 92); as discussed in section VI.4, in fact we can generally give characterizations of mathematical structures that can be satisfied by systems of real world objects and are independent of general mathematical proof systems like set theory. One of Maddy's main priorities is advocating for the importance of extrinsic evidence in mathematics (ibid., Chapter V). As discussed in section VI.7 however, extrinsic evidence does not appear to be very useful for directly arguing for the soundness of a

proof system – and thus for its status as a candidate foundation. If one is convinced by the iterative conception of sets, or the limitation of size conception, then it is possible that one could try to confirm which further, stronger axioms hold of the set theoretic hierarchy by considering those axioms' consequences – the kind of confirmation Maddy is most concerned with (ibid., Chapters II, V). However the problem is that one is then trying to use extrinsic evidence from the point of view of a a robust realist – in Maddy's terms – and as she discusses, there seems to be no reason why axioms that merely have attractive consequences in this way should turn out to be true, in fact: some sort of further argument linking attractive consequences with truth would seem to be required (ibid., pp. 57–58).

It is this worry that leads Maddy to develop her thin realist and arealist viewpoints on set theory. The idea behind the former is that sets just are the things that set theoretic methods track (ibid., p. 61), so that set theoretic methods are automatically accurate ways of reasoning about them. On the latter viewpoint, even this thin form of existence is jettisoned (ibid., pp. 88–89), and set theory is viewed as a field without entities for its subject matter, but still with a form of objectivity governed by mathematical depth. Either way, the idea is to avoid the hard questions about why the extrinsic evidence Maddy focuses on should actually indicate the truth of certain set theoretic axioms. When faced with the question of soundness, there is no way to avoid these hard questions though: one cannot assume that these set theoretic methods – justifying axioms by extrinsic evidence in this way – automatically track the truths of set theory, since one cannot assume that these set theoretic methods justify the sound axiom systems in particular. It is perfectly plausible for instance that one could justify ZFC + MC or ZFC+CH by extrinsic evidence of this kind, but that in fact ZFC+$(V = L)$ or ZFC+$\neg$CH could turn out to be the sound theories. This possibility cannot be dismissed *a priori* by the kind of manoeuvres Maddy hopes to carry out.

Though the arguments of this chapter are intended to establish the need for sound-

ness largely by a consideration of what is important to our mathematics, and what we expect from mathematical results – the kinds of questions that philosophers of mathematical practice might consider – we are led in this way back to the questions of justification and truth, as the truth of either the iterative conception or the limitation of size conception of set seems to be the only plausible candidate for justifying the soundness of our current foundation. Thus we end up with a view of foundations more like that of Linnebo and Pettigrew (2011) and Ladyman and Presnell (2018) than the other authors considered above, though there are still differences. Linnebo and Pettigrew (2011) are concerned with the question of whether ETCS forms an autonomous foundation for set theory, and as discussed in section VI.7, there is no need for autonomy when it comes to soundness – it is perfectly plausible that we could justify the soundness of ETCS by first justifying the soundness of a membership based theory such as ZFC or ZU. Ladyman and Presnell (2018) give an account of foundations in which a foundation may consist of five components: a framework for mathematics, a semantics for this framework, a metaphysics for this semantics, an epistemology for this metaphysics, and a methodology for mathematical practice. From the point of view here, the first is essential, and the fifth arises from the first as discussed in chapter I. The second, third and fourth may all be important, in as much as these questions contribute to establishing the soundness of a proof system, as they do for ZFC.

Overall, this chapter can be seen as making a case for *veritism*, which we can define (in the context of philosophy of mathematics at least) as the view that it matters whether the principles of the proof system we use in mathematics are true. Indeed, according to the arguments of this chapter, unless the iterative conception of set or the limitation of size conception do establish that the axioms of ZFC(U) hold collectively of the sets that exist – or some other account which establishes the truth of these axioms is available – we should not be using ZFC as our foundation.

Conclusion

The focus of this thesis has been on mathematical proof: how it is done, and how it should be done. Chapter I started by giving a new account of what the standard of rigour in mathematics consists in, discussing some attractive properties of this standard, and defending the account from some objections. Chapters II and III then considered contrasting views of proof based around the result from knot theory known as Alexander's lemma, with chapter II undermining De Toffoli and Giardino's comments about this result, and diffusing the threat their view of proof poses to the kind of account given in chapter I, and chapter III arguing that Jones's comments about the argument also pose no threat to that account of rigour, using his version of the argument to illustrate that account, and also to illustrate general conditions that rigour requires of pictorial arguments.

Then chapter IV and chapter V discussed conceptual and logical issues that are relevant to the discussion of soundness in chapter VI: with chapter IV discussing the basis of primitive recursion, arguing for a view of it as founded on a logical operator called the double ancestral, and incidentally using this to strengthen an argument for Isaacson's thesis; and chapter V combining the ancestral and double ancestral operators with plural logic, to give a new account of the concepts of finiteness and equinumerosity for finite pluralities, and thus a new interpretation of arithmetic, contrasted in places with the Neo-Fregean interpretation.

Finally with this work done, chapter VI gave an argument for what we should require

of a proof system for it to be a candidate foundation for mathematics: namely, that it be sound for as many kinds of structures as possible, i.e. that for as many kinds of structures as possible, when a generalization about that kind of structure is provable in the proof system, that generalization actually holds of all real examples of that kind of structure which exist. It was argued that our best prospects of arguing for soundness are via the iterative conception of set, or the limitation of size conception; that the view of set theory of Maddy (2011), that extrinsic justifications are key, and automatically accurate, is thus mistaken; and that – if the arguments of this thesis are correct – our use of ZFC as a foundation is only appropriate if some such conception of the totality of sets is valid, and does justify belief in the axioms of the proof system.

Appendices

Appendix A

The Smooth Case of Alexander's Lemma

This appendix supplies the mathematical details for the rigorous reconstruction of Jones's argument, discussed in chapter III. The reconstructed proof is found in section A.1. The proof refers to basic facts about smooth and periodic functions, and smooth knots, which are collected in section A.2 and section A.3 respectively. An effort is made to use elementary arguments, or arguments as elementary as can be hoped, though there is one appeal to Sard's theorem in proposition A.3.23.

1 The proof

First, some notation. An interval is a subset $I \subseteq \mathbb{R}$ such that if $a, b \in I$ and $a \leqslant c \leqslant b$ then $c \in I$. A subset of $\mathbb{R}$ is thus an interval iff it is connected. The empty set is included by this definition as an interval. We say that an interval is **proper** if it contains at least two points, and is thus uncountable.

We write $\mathrm{Int}(A)$ for the interior of a subset A of a topological space. We write for instance $A + B$ for $\{a + b \mid a \in A, b \in B\}$ if A and B are subsets of some vector space,

and similarly write for instance λA for $\{\lambda a \mid a \in A\}$ if λ is a scalar.

If $T > 0$, a T-periodic function is a function f with domain $\mathbb{R}$ such that $f(x+T) = f(x)$ for all x. From now on we fix $T > 0$ to serve as the period of our periodic functions (and in particular, our knots). If V is a subspace of $\mathbb{R}^n$, we let $C^\infty_T(\mathbb{R}, V)$ denote the space of smooth T-periodic functions from $\mathbb{R}$ to V.

For $t, t' \in \mathbb{R}$, we write $t \equiv t'$ if there is $k \in \mathbb{Z}$ with $t' - t = Tk$, i.e. if $t' - t \in T\mathbb{Z}$. This is an equivalence relation on $\mathbb{R}$. We let $\frac{\mathbb{R}}{T\mathbb{Z}}$ be the quotient of $\mathbb{R}$ by $\equiv$, and let $\pi : \mathbb{R} \to \frac{\mathbb{R}}{T\mathbb{Z}}$ be the quotient map. $\frac{\mathbb{R}}{T\mathbb{Z}}$ is compact since it is the image of $[0, T]$, and π is open since if $U \subseteq \mathbb{R}$ then $\pi^{-1}(\pi(U)) = U + T\mathbb{Z} = \bigcup_{k\in\mathbb{Z}} U + k$ is open, so by the definition of the quotient topology $\pi(U)$ is open.

Now for the definition of smooth knot.

Definition 1.1. A **smooth knot** is a smooth map $\gamma : \mathbb{R} \to \mathbb{R}^3$ with period T such that for all t, $\gamma'(t) \neq 0$, and such that γ is injective up to $\equiv$-equivalence – i.e. for any $t, t' \in \mathbb{R}$ we have $\gamma(t) = \gamma(t')$ iff $t \equiv t'$.

In general, if A is a proper interval in $\mathbb{R}$ we call a smooth curve $\beta : A \to \mathbb{R}^n$ an **immersion** if for all t, $\beta'(t) \neq 0$. We say that a pair $(s, t) \in \mathbb{R}^2$ is a **crossing pair** of $\beta \in C^\infty_T(\mathbb{R}, \mathbb{R}^n)$ if $s \not\equiv t$ and $\beta(s) = \beta(t)$. We say that a point $s \in \mathbb{R}$ is a **crossing point** of β if there is some t such that (s, t) is a crossing pair of β. Thus smooth knots are T-periodic smooth immersions with no crossing pairs, and no crossing points.

Definition 1.2. Let β and γ be smooth knots in $\mathbb{R}^3$. A **smooth isotopy** from β to γ is a smooth map $H : [0, 1] \times \mathbb{R} \to \mathbb{R}^3$ such that if we set $H_s : t \mapsto H(s, t)$ then H_s is a smooth knot for all $s \in [0, 1]$, and $H_0 = \beta$ and $H_1 = \gamma$. We write $\beta \sim \gamma$ if a smooth isotopy from β to γ exists.

As shown in proposition A.3.1, this relation $\sim$ is an equivalence relation.

When working with knots it is helpful to have a notion of when two curves in space

are "close". To this end if V is a subspace of $\mathbb{R}^n$ we equip $C^\infty_T(\mathbb{R}, V)$ with the norm

$$\|\gamma\|_{C^1} = \sup_{t\in A}\|\gamma(t)\| + \sup_{t\in A}\|\gamma'(t)\|.$$

An important fact about smooth knots is that if γ is a smooth knot, and β is any T-periodic smooth curve that is close enough to γ, then β is a smooth knot which is smoothly isotopic to γ (theorem A.3.8).

Let P be a plane in $\mathbb{R}^3$. Any point $x \in \mathbb{R}^3$ can be written uniquely as $x = p_x + q_x$ with $p_x \in P$ and q_x orthogonal to P. The map $x \mapsto p_x$ is called the **projection onto** $\boldsymbol{P}$. It is an affine linear and therefore smooth map, and we denote it by π_P. If γ is a curve, we write γ_P for the curve $\pi_P \circ \gamma$.

We will focus our attention on the plane $\{(x, y, 0) \mid x, y \in \mathbb{R}\}$, which we identify with $\mathbb{R}^2$ and with $\mathbb{C}$. Questions about projection onto any other plane can be reduced to questions about projection onto $\mathbb{C}$ by a suitable rotation (and translation) of space.

Recall that a crossing pair of a T-periodic curve β is a pair (s, t) such that $s \not\equiv t$ and $\beta(s) = \beta(t)$, and that a crossing point is a point s such that (s, t) is a crossing pair for some t.

Definition 1.3. Let $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{C})$. Say that γ is **regular** in $\mathbb{C}$ if:

(i) Crossing points of γ are only double crossings, so that if s is a crossing point of γ then $|\{t \in [0, T) \mid \gamma(t) = \gamma(s)\}| = 2$

(ii) γ is an immersion, so for all t, $\gamma'(t) \neq 0$

(iii) Crossings are transversal, so that if (s, t) is a crossing pair of γ, then $\gamma'(s)$ and $\gamma'(t)$ are not parallel

If $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{R}^3)$, say that γ has **regular projection** if $\gamma_{\mathbb{C}}$ is regular.

If γ has regular projection then theorem A.3.10 shows that crossing points form a

closed and isolated set, as do crossing pairs, and there are only finitely many $t \in [0, T)$ such that $\gamma_{\mathbb{C}}$ has a crossing point at t.

The following is basic and crucial.

Theorem 1.4. *The set of smooth knots which have regular projections is open and dense in the set of smooth knots.*

Proof. We have that $C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{C})$ is open by proposition A.3.2 and dense by proposition A.3.20. Then if we let G be the set of $\gamma \in C_T^{\infty}(\mathbb{R}, \mathbb{C})$ which are regular, then G is an open subset of $C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{C})$ by theorem A.3.17 and thus an open subset of $C_T^{\infty}(\mathbb{R}, \mathbb{C})$. Also G is a dense subset of $C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{C})$ by theorem A.3.26 and thus a dense subset of $C_T^{\infty}(\mathbb{R}, \mathbb{C})$.

Now let S be the set of smooth knots, K the set of smooth knots which have regular projections. If we let

$$\mathrm{proj} : C_T^{\infty}(\mathbb{R}, \mathbb{R}^3) \to C_T^{\infty}(\mathbb{R}, \mathbb{C}), \quad \gamma \mapsto \gamma_{\mathbb{C}}$$

then we have that K is $(\mathrm{proj}|_S)^{-1}(G)$. By proposition A.3.18 proj is continuous, so K is an open subset of S. Then we claim that $\mathrm{proj}^{-1}(G)$ is dense in $C_T^{\infty}(\mathbb{R}, \mathbb{R}^3)$. Indeed let $\gamma \in C_T^{\infty}(\mathbb{R}, \mathbb{R}^3)$ and $\epsilon > 0$. Then we can find $\alpha \in G$ with $\|\alpha - \gamma_{\mathbb{C}}\|_{C^1} < \epsilon$. Then if we let $\beta = \gamma + (\alpha - \gamma_{\mathbb{C}})$ we have $\beta_{\mathbb{C}} = \alpha$ and so $\beta \in \mathrm{proj}^{-1}(G)$, but also $\|\beta - \gamma\|_{C^1} = \|\alpha - \gamma_{\mathbb{C}}\|_{C^1} < \epsilon$ as required. Thus $\mathrm{proj}^{-1}(G)$ is dense in $C_T^{\infty}(\mathbb{R}, \mathbb{R}^3)$, so $K = (\mathrm{proj}|_S)^{-1}(G) = S \cap \mathrm{proj}^{-1}(G)$ is dense in S, since S is open. Thus K is both open and dense in S, as required. □

To discuss Alexander's lemma, we need a way of talking about "which direction a path goes around p_0" where p_0 is a point in our plane $\mathbb{C}$. For simplicity, we will just cover the case $p_0 = 0 \in \mathbb{C}$, for which the existing notion of complex argument suffices.

If γ is a curve in $\mathbb{R}^3$ such that $0 \notin \mathrm{Image}(\gamma_{\mathbb{C}})$ then we say that γ projects avoiding 0.

Proposition 1.5. *The set of smooth knots γ which have regular projection onto $\mathbb{C}$ position avoiding 0 is open and dense in the set of smooth knots.*

Proof. Immediate by combining proposition A.3.19 with theorem A.1.4. □

Thus any smooth knot β is smoothly isotopic to such a γ, by theorem A.3.8.

Now suppose γ is a smooth knot which projects avoiding 0. $\exp : \mathbb{C} \to \mathbb{C}^\times$ is the smooth universal cover of $\mathbb{C}^\times$, so there is a smooth lifting $\tilde{\gamma} : \mathbb{R} \to \mathbb{C}$ such that $\exp \circ \tilde{\gamma} = \gamma_{\mathbb{C}}$, and $\tilde{\gamma}$ is unique up to a translation of the plane by $2k\pi i$, $k \in \mathbb{Z}$. We will call such a $\tilde{\gamma}$ a **lifting of $\gamma_{\mathbb{C}}$ through exp**. We will use the notation Arg_γ for the second component of $\tilde{\gamma}$, which is appropriate since for every t we have $\gamma(t) = e^{\tilde{\gamma}_1(t)} e^{i\mathrm{Arg}_\gamma(t)}$. We call such a function Arg_γ a **smooth argument function for γ**.

Note that we have

$$\gamma'_{\mathbb{C}}(t) = \tilde{\gamma}'(t) \exp(\tilde{\gamma}(t)) = \tilde{\gamma}'(t)\gamma_{\mathbb{C}}(t)$$

so since $\gamma_{\mathbb{C}}(t) \neq 0$ we have that $\gamma'_{\mathbb{C}}(t) = 0$ iff $\tilde{\gamma}'(t) = 0$. Thus $\gamma_{\mathbb{C}}$ is an immersion iff $\tilde{\gamma}$ is.

We similarly write for instance $\tilde{\gamma}$ for a smooth lifting through exp of γ if γ is a curve in $\mathbb{C}^\times$.

The above gives us the ability to talk about "the direction $\gamma_{\mathbb{C}}$ is going around 0 at t". Letting $D_\gamma(t) = \mathrm{Arg}'_\gamma(t)$, we have

$$\{t \mid (\mathrm{Arg}(\gamma_{\mathbb{C}}))'(t) > 0\} = \{t \mid D_\gamma(t) > 0\}$$

being the set of t at which $\gamma_{\mathbb{C}}$ is going anti clockwise around 0, and

$$\{t \mid (\mathrm{Arg}(\gamma_{\mathbb{C}}))'(t) < 0\} = \{t \mid D_\gamma(t) < 0\}$$

being the set of t at which $\gamma_{\mathbb{C}}$ is going clockwise around 0. If $D_\gamma(t) = 0$ then $\gamma_{\mathbb{C}}$ is

going towards or away from 0 at t. Thus the function D_γ indicates which direction γ is going around zero. It is visibly smooth, and is independent of the choice of $\tilde{\gamma}$. It can be checked that this coincides with other ways of making rigorous the concept of which direction $\gamma_{\mathbb{C}}$ is going around 0.

Then the result we seek to prove is the following.

Alexander's lemma. *Let γ be a smooth knot. Then there is a smooth knot β which is smoothly isotopic to γ such that β has regular projection avoiding 0, and we have $D_\beta(t) > 0$ for all t.*

This is finally reached below as lemma A.1.24.

To prove this, we first split the knot up into sections on which it goes the wrong way. This is not quite so straightforward as in the polygonal case (Alexander 1923), as a section on which the knot goes the wrong way can smoothly bend at each end into sections on which it goes the right way.

We will define what we require of these sections below. First, some notation. If γ is a smooth knot which projects avoiding 0, and A is a subset of $\mathbb{R}$, we will use for instance the notation $D_\gamma^{\leqslant 0}(A)$ for $\{t \in A \mid D_\gamma(t) \leqslant 0\}$, a closed subset of A. We will similarly use the notation $D_\gamma^{>0}(A)$ for $\{t \in A \mid D_\gamma(t) < 0\}$ and so on (this one is an open subset of A).

Definition 1.6. Let I be a nonempty compact interval in $\mathbb{R}$. Let γ be a smooth knot which projects avoiding 0. We say that γ has **at most one backwards bend** on I if $\sup(I) < \inf(I) + \frac{T}{2}$, and $D_\gamma^{\leqslant 0}(I)$ is an interval such that

$$\sup\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(I)\} \leqslant \inf\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(I)\} + \pi.$$

Then if J is any nonempty bounded interval we say that γ has **at most one backwards bend** on J if it has at most one backwards bend on $\overline{J}$.

This is independent of the choice of Arg_γ. We require $\sup(I) < \inf(I) + \frac{T}{2}$ so that two such sections can only intersect at one end modulo $\equiv$, which simplifies things a little.

The condition

$$\sup\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(I)\} \leqslant \inf\{\mathrm{Arg}_\gamma(t) \mid t \in D_\gamma^{\leqslant 0}(I)\} + \pi.$$

controls the argument of γ on $D_\gamma^{\leqslant 0}(\overline{I})$, so that γ always stays in some half plane through the origin. It follows from this that

$$\mathrm{Arg}_\gamma(\inf D_\gamma^{\leqslant 0}(I)) \leqslant \mathrm{Arg}_\gamma(\sup D_\gamma^{\leqslant 0}(I)) + \pi,$$

and this is actually an equivalent statement, since Arg_γ is decreasing on $D_\gamma^{\leqslant 0}(I)$ (since $\mathrm{Arg}'_\gamma(t) = D_\gamma(t)$).

Note that the cases where I is a singleton, or where $D_\gamma^{\leqslant 0}(I)$ is empty, are allowed under this definition. Note also that if γ has at most one backwards bend on I, and J is a nonempty interval with $J \subseteq I$ then γ has at most one backwards bend on J. Note finally that if $\sup(I) < \inf(I) + \frac{T}{2}$ and $D_\gamma(t) > 0$ for all $t \in \overline{I}$ then γ has at most one backwards bend on I.

When folding parts of the knot back around itself the results will be of the following form.

Definition 1.7. Let γ be a smooth knot which has regular projection avoiding 0. Let I be a compact interval with $\sup(I) - \inf(I) < \frac{T}{2}$. Say that a smooth knot β is a **bending forwards of γ on** I if it has regular projection avoiding 0, is smoothly isotopic to γ and:

(i) $D_\beta^{>0}(I)$ is an interval

(ii) If $t \notin \mathrm{Int}(I) + T\mathbb{Z}$ then $\beta(t) = \gamma(t)$

(iii) $D_\beta(t) \geqslant D_\gamma(t)$ for all t

A basic fact about bending forwards is that it is transitive in a sense.

Proposition 1.8. *Suppose β is a bending forwards of γ on I and α is a bending forwards of β on J, such that there is $t \in I \cap J$ with $D_\beta(t) > 0$ and such that $\sup(I \cup J) - \inf(I \cup J) < \frac{T}{2}$. Then α is a bending forwards of γ on $I \cup J$.*

Proof. We have that α has regular projection avoiding 0, and is smoothly isotopic to γ.

If $t \in D_\alpha^{>0}(I \cup J)$ and $t \notin J$ then $D_\alpha(t) = D_\beta(t)$ so $D_\beta(t) > 0$, so $t \in D_\beta^{>0}(I)$. In other words $D_\alpha^{>0}(I \cup J) = D_\alpha^{>0}(J) \cup D_\beta^{>0}(I)$, which are both intervals, with nonempty intersection since there is $t \in I \cup J$ with $D_\beta(t) > 0$. Thus $D_\alpha^{>0}(I \cup J) = D_\alpha^{>0}(I) \cup D_\beta^{>0}(J)$ is an interval.

Also we have that if $t \notin \mathrm{Int}(I \cup J) + T\mathbb{Z}$ then $t \notin \mathrm{Int}(I) + T\mathbb{Z}$ and $t \notin \mathrm{Int}(J) + T\mathbb{Z}$, so $\alpha(t) = \beta(t) = \gamma(t)$. Finally for all t we have $D_\alpha(t) \geqslant D_\beta(t) \geqslant D_\gamma(t)$. □

The reason for modifying γ according to these constraints is the following.

Proposition 1.9 (Bending forwards does not interfere with other sections)**.** *Suppose γ is a smooth knot, and I is a compact interval such that γ has at most one backwards bend on I. Suppose that β is a bending forwards of γ on J such that that $D_\beta^{>0}(J) \nsubseteq \mathrm{Int}(D_\gamma^{\leqslant 0}(I)) + T\mathbb{Z}$. Then β has at most one backwards bend on I.*

Proof. First, let $W = I + T\mathbb{Z}$. If $W \cap J = \varnothing$ then $\beta|_I \gamma|_I$ so β has at most one backwards bend on I and we are done.

Note that there is at most one $k \in \mathbb{Z}$ such that $I + kT \cap J \neq \varnothing$. If we have $u + kT \in J$ and $v + lT \in J$ with $u, v \in I$ and $k, l \in \mathbb{Z}$ then we have $|(u + kT) - (w + lT)| \leqslant \sup(J) - \inf(J) < \frac{T}{2}$, so

$$\begin{aligned} |k - l|T &\leqslant |(u - w) + (k - l)T| + |u - w| \\ &= |(u + kT) - (w + lT)| + |u - w| < \frac{T}{2} + \frac{T}{2} = T, \end{aligned}$$

so $k = l$.

Thus we can assume that there is exactly one $k \in \mathbb{Z}$ with $(I + kT) \cap J \neq \varnothing$. Since γ and β are periodic, if $k \in \mathbb{Z}$ then the proposition holds for I iff it holds for $I + kT$. Thus we can assume $I \cap J \neq \varnothing$. Then if $t \in I$ and $t \notin \mathrm{Int}(J)$ then $t \notin \mathrm{Int}(J) + T\mathbb{Z}$, so $\beta(t) = \gamma(t)$ and $D_\beta(t) = D_\gamma(t)$.

We claim that $D_\beta^{\leqslant 0}(I) = D_\gamma^{\leqslant 0}(I) \setminus D_\beta^{>0}(J)$. Certainly if $t \in D_\beta^{\leqslant 0}(I)$ then $D_\gamma(t) \leqslant D_\beta(t) \leqslant 0$, so $t \in D_\gamma^{\leqslant 0}(I)$, and necessarily $t \notin D_\beta^{>0}(J)$. Conversely if $t \in D_\gamma^{\leqslant 0}(I) \setminus D_\beta^{>0}(J)$ then if $t \notin J$ we have $D_\beta(t) = D_\gamma(t) \leqslant 0$, and if $t \in J$ then we have $t \in J \setminus D_\beta^{>0}(J)$ so $D_\beta(t) \leqslant 0$, as required.

Now we argue that $D_\gamma^{\leqslant 0}(I) \setminus D_\beta^{>0}(J)$ is an interval, i.e. that if $a, b \in D_\gamma^{\leqslant 0}(I) \setminus D_\beta^{>0}(J)$ and $a \leqslant c \leqslant b$ then $c \in D_\gamma^{\leqslant 0}(I) \setminus D_\beta^{>0}(J)$. But since $D_\gamma^{\leqslant 0}(I)$ is an interval, the only way this could fail is if $a, b \in D_\gamma^{\leqslant 0}(I) \setminus D_\beta^{>0}(J)$ and $c \in D_\beta^{>0}(J)$. If this were the case then since $D_\beta^{>0}(J)$ is an interval we would have $D_\beta^{>0}(J) \subseteq (a, b) \subseteq \mathrm{Int}(D_\gamma^{\leqslant 0}(I))$, contradicting a premise. Thus we do indeed have that $D_\gamma^{\leqslant 0}(I) \setminus D_\beta^{>0}(J) = D_\beta^{\leqslant 0}(I)$ is an interval.

Finally we need to control the angle of β on $D_\beta^{\leqslant 0}(I)$. If $t, t' \in D_\gamma^{\leqslant 0}(I)$ we have

$$\mathrm{Arg}_\gamma(t) \leqslant \mathrm{Arg}_\gamma(t') + \pi.$$

Let $\tilde{\beta}$ be a lifting of β through exp. We have $0 \geqslant D_\beta(t) \geqslant D_\gamma(t)$ for $t \in D_\beta^{\leqslant 0}(I)$, i.e. $\mathrm{Arg}'_\gamma(t) \leqslant \mathrm{Arg}'_\beta(t) \leqslant 0$, so if $t \leqslant t' \in D_\beta^{\leqslant 0}(I)$ then $\mathrm{Arg}_\gamma(t') - \mathrm{Arg}_\gamma(t) \leqslant \mathrm{Arg}_\beta(t') - \mathrm{Arg}_\beta(t) \leqslant 0$. Thus

$$-\pi \leqslant \mathrm{Arg}_\gamma(t') - \mathrm{Arg}_\gamma(t) \leqslant \mathrm{Arg}_\beta(t') - \mathrm{Arg}_\beta(t) \leqslant 0$$

so that indeed

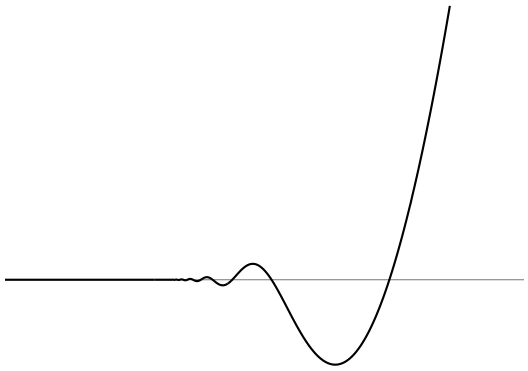
$$\sup\{\mathrm{Arg}_\beta(t) \mid t \in D_\beta^{\leqslant 0}(I)\} \leqslant \inf\{\mathrm{Arg}_\beta(t) \mid t \in D_\beta^{\leqslant 0}(I)\} + \pi. \qquad \square$$

Let γ be a smooth knot which has regular projection avoiding 0. The proof strategy is

to cover the stretches where $\gamma_{\mathbb{C}}$ bends the wrong way with finitely many open intervals on which it has at most one backwards bend (as defined above), and then go through these intervals one by one correcting them. One problem with this is that some troublesome stretches of $\gamma_{\mathbb{C}}$ may not be covered by intervals on which γ has at most one backwards bend. Smooth functions can be very wriggly. For instance the function

$$f(x) = \begin{cases} 0 & \text{if } x \leqslant 0 \\ \sin\left(\frac{1}{x}\right)e^{-\frac{1}{x}} & \text{if } x > 0 \end{cases}$$

is smooth, and wiggles above and below the y-axis infinitely many times as x approaches 0 from above:



Similarly there may be a point $t \in \mathbb{R}$ such that in every neighbourhood of t, D_γ oscillates above and below zero infinitely many times. Thus t cannot be contained in a neighbourhood on which γ has at most one backwards bend, as defined above.

This factor complicates the argument. One potential response might be to observe that if D_γ oscillates above and below zero near t then the diagram of γ is roughly radial near t, roughly heading straight out from or straight towards the axis, and to try to argue that these roughly radial sections can also be bent forwards (as sections on which γ has at most one backwards bend can be). But filling in the details of this would mean

much extra work. We would need a precise definition of what we mean by "roughly straight" here, paralleling that of "at most one backwards bend" given above; we will need an analogue of proposition A.1.9 for these roughly straight sections; and we will also (probably) need to give two separate arguments justifying the over the shoulder manoeuvre – one for sections with at most one backwards bend, and one for roughly straight sections.

Fortunately, all this can be avoided with a small change to γ. Note that if $D_\gamma(t) = 0$ but $D'_\gamma(t) \neq 0$ then the zero of D_γ is isolated – there is a neighbourhood U of t such that t is the only point of U at which D_γ is zero. Such points are called simple zeroes of D_γ. If D_γ has only simple zeroes then $\{t \mid D_\gamma(t) = 0\}$ is a closed isolated set, so has finite intersection with $[0, T]$.

Though it might be possible to "by hand", going through all the points on the diagram of γ where D_γ does not have a simple zero and adjusting them appropriately, that looks like it could well turn out to be a lot of work. Instead we can obtain the result by adding a small perturbation to the diagram of γ. By considering all small perturbations of a certain form, we can argue that one will have the desired result.

Proposition 1.10. *The set of smooth knots γ which project avoiding 0 such that D_γ has only simple zeroes is dense in the set of smooth knots.*

Proof. Let γ be a smooth knot which projects onto $\mathbb{C}^\times$ and let $\epsilon > 0$. We may assume that ϵ is small enough that $B_\epsilon(\gamma)$ consists only of smooth knots β which project onto $\mathbb{C}^\times$, by proposition A.3.19. Let $\tilde{\gamma}$ be a lifting of γ through exp. By proposition A.2.10, there is $\delta > 0$ such that if $\|\tilde{\beta} - \tilde{\gamma}\|_{C^1} < \delta$ then $\|\exp \circ \tilde{\beta} - \gamma\|_{C^1} < \epsilon$.

Let

$$\alpha : \mathbb{R} \to \mathbb{C},\ t \mapsto \left(-\frac{T}{2\pi}\mathrm{Arg}'_\gamma(t) + \left(\frac{T}{2\pi}\right)^2 \mathrm{Arg}''_\gamma(t)i \right) \cdot e^{-\frac{2\pi i t}{T}}.$$

This is a smooth curve. Thus by proposition A.2.9 we can find z such that $\|z\|(1+\frac{2\pi}{T}) < \delta$ and $z \notin \mathrm{Image}(\alpha)$.

Let $z = re^{i\theta}$, and let

$$\tilde{\beta} : \mathbb{R} \to \mathbb{C},\ t \mapsto \tilde{\gamma}(t) + r \sin\left(\frac{2\pi t}{T} + \theta\right) e_2.$$

This is periodic with period T. We have

$$\tilde{\beta}'(t) = \tilde{\gamma}'(t) + r\frac{2\pi}{T} \cos\left(\frac{2\pi t}{T} + \theta\right) e_2.$$

Then we have $\|\tilde{\beta}(t) - \tilde{\gamma}(t)\| \leqslant \|z\|$, and $\|\tilde{\beta}'(t) - \tilde{\gamma}'(t)\| \leqslant \|z\|(\frac{2\pi}{T})$. Thus $\|\tilde{\beta} - \tilde{\gamma}\|_{C^1} \leqslant \|z\|(1 + (\frac{2\pi}{T})) < \delta$, so $\|\exp \circ \tilde{\beta} - \gamma\|_{C^1} < \epsilon$. Thus if $D_{\exp \circ \tilde{\beta}}$ has only simple zeroes then we are done, i.e. if Arg'_β has only simple zeroes.

But if there is t such that $\mathrm{Arg}'_\beta(t) = \mathrm{Arg}''_\beta(t) = 0$, then we have

$$\mathrm{Arg}'_\gamma(t) + r\frac{2\pi}{T} \cos\left(\frac{2\pi t}{T} + \theta\right) = 0$$
$$\mathrm{Arg}''_\gamma(t) - r\left(\frac{2\pi}{T}\right)^2 \sin\left(\frac{2\pi t}{T} + \theta\right) = 0$$

so

$$-\frac{T}{2\pi}\mathrm{Arg}'_\gamma(t) = r \cos\left(\frac{2\pi t}{T} + \theta\right)$$
$$\left(\frac{T}{2\pi}\right)^2 \mathrm{Arg}''_\gamma(t) = r \sin\left(\frac{2\pi t}{T} + \theta\right)$$

or in other words

$$-\frac{T}{2\pi}\mathrm{Arg}'_\gamma(t) + \left(\frac{T}{2\pi}\right)^2 \mathrm{Arg}''_\gamma(t)i = re^{i\theta}e^{\frac{2\pi i t}{T}} = ze^{\frac{2\pi i t}{T}}$$

which contradicts the choice of z. □

The set of smooth knots γ with $\mathrm{Image}(\gamma_{\mathbb{C}}) \subseteq \mathbb{C}^\times$ such that D_γ has only simple zeroes

is not open in the set of smooth knots, under the topology we have given it, so it may be a bad idea to think of such knots as being in "general position". It looks like it would be open in the Whitney C^2-topology, but using that would complicate other arguments.

For brevity we will say that a smooth knot γ **projects nicely** if it has regular projection avoiding 0, with D_γ only having simple zeroes. By proposition A.1.10 together with proposition A.1.5 we have that the set of smooth knots which project nicely is dense in the set of smooth knots. Thus by theorem A.3.8, every smooth knot is isotopic to a smooth knot which projects nicely. Thus it suffices to prove Alexander's lemma for smooth knots which project nicely.

If γ projects nicely then every troublesome part of $\gamma_{\mathbb{C}}$ can be covered by an open interval on which γ bends backwards. Indeed, if t is a point such that $D_\gamma(t) \leqslant 0$ then either $D_\gamma(t) < 0$ or $D_\gamma(t) = 0$. In the former case, t is contained in an open interval U such that $D_\gamma < 0$ on $\overline{U}$, so that γ has at most one backwards bend on U. In the latter case, since D_γ has only simple zeroes, D_γ changes sign at t so there is an open interval U containing t such that $\{t' \in \overline{U} \mid D_\gamma(t') \leqslant 0\}$ is either $\{t' \in \overline{U} \mid t' \geqslant t\}$ or $\{t' \in \overline{U} \mid t' \leqslant t\}$.

Thus we have that $D_\gamma^{\leqslant 0}([0,T])$ is a compact set which is covered by intervals on which γ has at most one backwards bend, so is covered by finitely many such intervals. This gives us our proof strategy: go through these intervals, bending them the right way one by one. Indeed, having done this preliminary work, Alexander's lemma can be derived fairly straightforwardly from the following.

Bending Forward proposition. *Let γ be a smooth knot which projects nicely. Let U be an open interval in $\mathbb{R}$ such that γ has at most one backwards bend on U. Then there is a smooth knot β which projects nicely and is a bending forwards of γ such that if $t \in U$ then $D_\beta(t) > 0$.*

This is proved below as proposition A.1.21, and the proof of Alexander's lemma follows as lemma A.1.24.

To prove this Bending Forward proposition result we use the "over the shoulder" manoeuvre that Jones describes.

We start with the ability to push our section of knot into the plane, or pull it out away from the plane (depending on whether our stretch lies above or below the stretch it crosses).

Proposition 1.11. *Let γ be a smooth knot such that $\gamma'_{\mathbb{C}}(t) \neq 0$ for all t, and $\gamma_{\mathbb{C}}$ has at most one crossing point in (a, b), with $a < b < a + T$. Let $K > 0$ such that $|\gamma_3(t)| < K$ for all t, and let $a < c < e < f < d < b$. Then γ is smoothly isotopic to a smooth knot β with:*

- $\beta_{\mathbb{C}} = \gamma_{\mathbb{C}}$
- $\beta(t) = \gamma(t)$ *for* $t \notin (a, b) + T\mathbb{Z}$
- $|\beta_3(t)| \leqslant K$ *for* $t \notin (c, d) + T\mathbb{Z}$
- $|\beta_3(t)| \geqslant K$ *for* $t \in [c, d]$
- $|\beta_3(t)| = 2K$ *for* $t \in [e, f]$
- $(|\beta_3|)'(t) > 0$ *for* $t \in (c, e)$
- $(|\beta_3|)'(t) < 0$ *for* $t \in (f, d)$

Proof. First, note that if (t, t') is a crossing pair of $\gamma_{\mathbb{C}}$ with $t \in (a, b)$ then (t', t) is a crossing pair of $\gamma_{\mathbb{C}}$ with $t' \neq t$ so we must have $t' \notin (a, b)$. We can let $u \in \{\pm 1\}$ such that if $\gamma_{\mathbb{C}}$ has a crossing pair (t, t') with $t \in (a, b)$ then $\gamma_3(t) = \gamma_3(t') + \lambda u$ with $\lambda > 0$, i.e. $u\gamma'_3(t) > u\gamma'_3(t')$.

Now, let $\phi : \mathbb{R} \to \mathbb{R}$ be a smooth function such that $\phi(t) = 2K$ for $t \in [e, f]$, $\phi(t) = K$ for $t \leqslant c$ or $t \geqslant d$, and $\phi'(t) > 0$ for $t \in (c, e)$, and $\phi'(t) < 0$ for $t \in (f, d)$. Then let $\psi : \mathbb{R} \to \mathbb{R}$ be a smooth function such that $0 \leqslant \psi(t) \leqslant 1$ for all t, $\psi(t) = 0$ for $t \leqslant a$ or

$t \geqslant b$ and $\psi(t) = 1$ for $t \in [c, d]$. Then let

$$\beta : [b - T, a + T] \to \mathbb{R}^3,\ t \mapsto \gamma_1(t)e_1 + \gamma_2(t)e_2 + (\gamma_3(t) + \psi(t)(u\phi(t) - \gamma_3(t)))e_3,$$

and extend it to a T-periodic smooth curve, using proposition A.2.8. We will show that this is a smooth knot smoothly isotopic to γ. Let

$$H : [0, 1] \times \mathbb{R} \to \mathbb{R}^3,\ (s, t) \mapsto \gamma(t) + s(\beta(t) - \gamma(t)).$$

This is a smooth map, and for every s we have $(H_s)_{\mathbb{C}} = \gamma_{\mathbb{C}}$. Thus for every s we have for all t that $(H_s)'_{\mathbb{C}}(t) = \gamma'_{\mathbb{C}}(t) \neq 0$, so H_s is a smooth immersion. Note also that $H_s(t) = \gamma(t)$ for $t \notin (a, b) + T\mathbb{Z}$, and that $u\beta_3(t) \geqslant u\gamma_3(t)$ for all t, so $u(H_s)_3(t) \geqslant u\gamma_3(t)$ for all s, t.

Now we show that each H_s has no crossing pairs. If $t, t' \notin (a, b) + T\mathbb{Z}$ then $H_s(t) = \gamma(t)$ and $H_s(t') = \gamma(t')$ so (t, t') is not a crossing pair of H_s. If (t, t') is a crossing pair of $(H_s)_{\mathbb{C}}$ with $t \in (a, b)$ then t is a crossing point of $\gamma_{\mathbb{C}}$, so is the unique crossing point of $\gamma_{\mathbb{C}}$ on (a, b), and we have $t' \notin (a, b)$ as noted at the start of the proof. But then we have $u(H_s)_3(t) \geqslant u\gamma_3(t) > u\gamma_3(t') = u(H_s)_3(t')$ so $(H_s)_3(t) \neq (H_s)_3(t')$ and so (t, t') is not a crossing pair of H_s. Thus H_s has no crossing pairs, so is a smooth knot. This holds for all s, so H is a smooth isotopy from γ to $H_1 = \beta$.

To conclude, note that $\beta_{\mathbb{C}} = \gamma_{\mathbb{C}}$, and $\beta(t) = \gamma(t)$ for $t \notin (a, b) + T\mathbb{Z}$. If $t \notin (c, d) + T\mathbb{Z}$ then $u\phi(t) \in [-K, K]$ and $\gamma_3(t) \in [-K, K]$, so $\beta_3(t) \in [-K, K]$. If $t \in [c, d]$ then then $|\beta_3(t)| = |\gamma_3(t) + u\phi(t) - \gamma_3(t)| = |u\phi(t)| = \phi(t)$, so $|\beta(t)| \geqslant K$, and we have $|\beta_3(t)| = 2K$ for $t \in [e, f]$ and $|\beta_3|'(t) > 0$ for $t \in (c, e)$ and $|\beta_3|'(t) < 0$ for $t \in (f, d)$. □

In order to suitably bend a piece of knot forwards, it is useful to be able to smoothly bend one planar curve into another while controlling the location of points on the curve. Let I be a proper interval in $\mathbb{R}$, let $n \in \mathbb{N}^\times$ and let $a \in I$ and $X \subseteq \mathbb{R}^n$. We say that a

smooth map $H : [0,1] \times I \to \mathbb{R}^n$ is a **smooth isotopy relative $(a \hookrightarrow X)$** if H is smooth, and for all s if we write $H_s : t \mapsto H(s,t)$ then H_s is an injective smooth immersion with $H_s(a) \in X$. We say such a map is a smooth isotopy relative $(a \hookrightarrow X)$ from H_0 to H_1. For any $a, b \in \mathbb{R}$, $X, Y \subseteq \mathbb{R}^n$, and injective smooth immersions $\alpha, \beta : I \to \mathbb{R}^n$, the relation of there being a smooth isotopy relative $(a \hookrightarrow X)$ and relative $(b \hookrightarrow Y)$ is an equivalence relation (by almost the same argument as proposition A.3.1).

Proposition 1.12. *Let I be a proper interval and let $\alpha, \beta : I \to \mathbb{R}^2$ be injective smooth immersions. Let $a \neq b \in I$ and let X, Y be disjoint convex subsets of $\mathbb{R}^2$ with $\alpha(a), \beta(a) \in X$ and $\alpha(b), \beta(b) \in Y$. Then there is a smooth isotopy relative $(a \hookrightarrow X)$ and relative $(b \hookrightarrow Y)$ from α to β.*

Proof. We will show that for any such α, a, b, X and Y, and any $x \in X$ and $y \in Y$, there is a smooth isotopy relative $(a \hookrightarrow X)$ and relative $(b \hookrightarrow Y)$ from α to the line

$$\lambda_{x,y} : t \mapsto \frac{t-a}{b-a} \cdot (y - x) + x.$$

Since the existence of a smooth isotopy relative $(a \hookrightarrow X)$ and relative $(b \hookrightarrow Y)$ is an equivalence relation, this will prove the proposition.

X is convex so the map $\sigma_a : [0,1] \to \mathbb{R}^2,\ s \mapsto x + s(\alpha(a) - x)$ is a smooth map into X, and similarly $\sigma_b : [0,1] \to \mathbb{R}^2,\ s \mapsto y + s(\alpha(b) - y)$ is a smooth map into Y. For all s, $\sigma_a(s) \neq \sigma_b(s)$ since X and Y are disjoint.

Let $\mu \in I$ and let

$$\psi : [0,1] \times I \to I,\ (s,t) \mapsto \mu + s(t - \mu).$$

This is smooth, and for each $s > 0$ the map $\phi_s : t \mapsto \phi(s,t)$ is a bijection from I to $\mu + s(I - \mu)$, with positive derivative. Thus for each $s > 0$, if we let $\alpha_s = \alpha \circ \phi_s$ then α_s is an injective smooth immersion.

We identify $\mathbb{R}^2$ with $\mathbb{C}$ and write $\cdot$ for complex multiplication on $\mathbb{C}$. For any $u, v, w, y \in \mathbb{C}$ with $u \neq v$ and $w \neq y$ the map $A_{u,v,w,y} : z \mapsto \frac{z-u}{v-u} \cdot (y-w) + w$ is a complex linear map which takes u to w and v to y. It is a combination affine rotation and dilation, with complex derivative $\frac{y-w}{v-u}$ which is non zero. Taking $A_s = A_{\alpha_s(a),\alpha_s(b),\sigma_a(s),\sigma_b(s)}$ for $s \neq 0$ we have that the map

$$A_s \circ \alpha_s : t \mapsto \frac{\alpha_s(t) - \alpha_s(a)}{\alpha_s(b) - \alpha_s(a)} \cdot (\sigma_b(s) - \sigma_a(s)) + \sigma_a(s)$$

is an injective smooth immersion, and we have for all s that $A_s(\alpha_s(a)) = \sigma_a(s) \in X$ and $A_s(\alpha_s(b)) = \sigma_b(s) \in Y$. We also have $A_1 = A_{\alpha_1(a),\alpha_1(b),\sigma_a(1),\sigma_b(1)} = A_{\alpha(a),\alpha(b),\alpha(a),\alpha(b)} = \mathrm{Id}$, so $A_1 \circ \alpha_1 = \alpha$. Thus to prove the proposition it suffices to show that the map

$$H : [0,1] \times I \to \mathbb{R}^2, (s,t) \mapsto \begin{cases} A_s(\alpha_s(t)) & \text{if } s \neq 0 \\ \frac{t-a}{b-a} \cdot (y-x) + x & \text{if } s = 0 \end{cases}$$

is smooth, as if it is smooth then it a smooth isotopy from $\lambda_{x,y}$ to α relative $(a \hookrightarrow X)$ and relative $(b \hookrightarrow Y)$. We have

$$H(s,t) = \begin{cases} \frac{\alpha_s(t)-\alpha_s(a)}{\alpha_s(b)-\alpha_s(a)} \cdot (\sigma_b(s) - \sigma_a(s)) + \sigma_a(s) & \text{if } s \neq 0 \\ \frac{t-a}{b-a} \cdot (\sigma_b(0) - \sigma_a(0)) + \sigma_a(0) & \text{if } s = 0 \end{cases}$$

so to prove that H is smooth it suffices to show that the function

$$J : [0,1] \times I \to \mathbb{R}^2,\ (s,t) \mapsto \begin{cases} \frac{\alpha_s(t)-\alpha_s(a)}{\alpha_s(b)-\alpha_s(a)} & \text{if } s \neq 0 \\ \frac{t-a}{b-a} & \text{if } s = 0 \end{cases}$$

is smooth. We have by proposition A.2.11 that the map

$$F_\alpha : I^2 \to \mathbb{R}^2,\ (t,t') \mapsto \begin{cases} \frac{\alpha(t)-\alpha(t')}{t-t'} & \text{if } t \neq t' \\ \alpha'(t) & \text{if } t = t' \end{cases}$$

is smooth. Thus the map

$$[0,1] \times I^2 \to \mathbb{R}^2,$$
$$(s,t,t') \mapsto \begin{cases} \frac{\alpha(\mu+s(t-\mu))-\alpha(\mu+s(t'-\mu))}{s(t-t')} & \text{if } s \neq 0 \text{ and } t \neq t' \\ \alpha'(\phi_s(t)) & \text{if } s \neq 0 \text{ and } t = t' \\ \alpha'(\mu) & \text{if } s = 0. \end{cases}$$

is smooth, so multiplying by $(t-t')$ we obtain that the map

$$f : [0,1] \times I^2 \to \mathbb{R}^2,$$
$$(s,t,t') \mapsto \begin{cases} \frac{\alpha_s(t)-\alpha_s(t')}{s} & \text{if } s \neq 0 \\ (t-t')\alpha'(\mu) & \text{if } s = 0 \end{cases}$$

is smooth, and non zero for all s and all $t \neq t'$. Thus we obtain a smooth map

$$[0,1] \times I \to \mathbb{R}^2,\ (s,t) \mapsto \frac{f(s,t,a)}{f(s,b,a)}$$
$$= \begin{cases} \left(\frac{\alpha_s(t)-\alpha_s(a)}{s}\right) \Big/ \left(\frac{\alpha_s(b)-\alpha_s(a)}{s}\right) & \text{if } s \neq 0 \\ \frac{(t-a)\alpha'(\mu)}{(b-a)\alpha'(\mu)} & \text{if } s = 0 \end{cases}$$
$$= \begin{cases} \frac{\alpha_s(t)-\alpha_s(a)}{\alpha_s(b)-\alpha_s(a)} & \text{if } s \neq 0 \\ \frac{(t-a)}{(b-a)} & \text{if } s = 0 \end{cases}$$
$$= J(s,t)$$

as required. □

Proposition 1.13. *Suppose $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{C})$ is regular with $a < b < a+T$ such that $\gamma(a) \neq \gamma(b)$, and $\beta \in C^\infty_T(\mathbb{R}, \mathbb{C})$ is a smooth immersion with $\beta(t) = \gamma(t)$ for $t \notin (a, b) + T\mathbb{Z}$. Then if*

$$J : [0, 1] \times \mathbb{R} \to \mathbb{R}^2,\ (s, t) \mapsto \gamma(t) + s(\beta(t) - \gamma(t)),$$

there is $h > 0$ such that for all s, $J_s|_{[b-T-h,\, a+h]}$ is a smooth immersion with its only crossing points on $[b - T,\, a]$.

Proof. By proposition A.3.21 there are $\delta, \epsilon > 0$ such that if $0 < h < \epsilon$ and $\alpha : [b - T - h, a + h] \to \mathbb{R}$ is smooth with $\alpha|_{[b-T,a]} = \gamma|_{[b-T,a]}$ and $\|\alpha'(t) - \gamma'(b - T)\| < \delta$ for $t \in [b - T - h, b - T]$ and $\|\alpha'(t) - \gamma'(a)\| < \delta$ for $t \in [d, d + h]$ then $\alpha|_{[b-T-h,a+h]}$ has no crossing points on $[b - T - h, b - T) \cup (a, a + h]$.

Since $\beta'(b - T) = \gamma'(b - T)$ and $\beta'(a) = \gamma'(a)$ we can find $h > 0$ such that $\|\beta'(t) - \gamma'(t)\| < \frac{\delta}{2}$ for $t \in [b - T - h, b] \cup [a, a + h]$. Then by shrinking h if necessary we may assume that $\|\gamma'(t) - \gamma'(b - T)\| < \frac{\delta}{2}$ for $t \in [b - T - h, b]$, and $\|\gamma'(t) - \gamma'(a)\| < \frac{\delta}{2}$ for $t \in [a, a + h]$. It follows that if $t \in [b - T - h, b]$ and $s \in [0, 1]$ then we have

$$\begin{aligned} \|J'_s(t) - \gamma'(b - T)\| &= \|\gamma'(t) + s(\beta'(t) - \gamma'(t)) - \gamma'(b - T)\| \\ &\leqslant \|\gamma'(t) - \gamma'(T - b)\| + s\|\gamma'(t) - \beta'(t)\| < \frac{\delta}{2} + \frac{s\delta}{2} \leqslant \delta. \end{aligned}$$

Similarly if $t \in [a, a+h]$ then $\|J'_s(t) - \gamma'(a)\| < \delta$. Thus for each $s \in [0, 1]$, J_s is a smooth immersion with no crossing points on $[b - T - h, b), (a, a + h]$. □

Proposition 1.14. *Let γ be a smooth knot which has regular projection, with $a < b < a + T$ such that $\gamma_{\mathbb{C}}(a) \neq \gamma_{\mathbb{C}}(b)$ and $\gamma_{\mathbb{C}}$ has at most one crossing point in (a, b). Suppose $\alpha \in C^\infty_T(\mathbb{R}, \mathbb{C})$ is a smooth immersion such that $\alpha|_{[a,b]}$ is injective, and with $\alpha(t) = \gamma_{\mathbb{C}}(t)$ for $t \notin (a, b) + T\mathbb{Z}$. Then γ is smoothly isotopic to a smooth knot β with $\beta(t) = \gamma(t)$ for $\notin (a, b) + T\mathbb{Z}$, and $\beta_{\mathbb{C}} = \alpha$.*

Proof. Let

$$J : [0,1] \times \mathbb{R} \to \mathbb{R}^2,\ (s,t) \mapsto \gamma_{\mathbb{C}}(t) + s(\alpha(t) - \gamma_{\mathbb{C}}(t)),$$

so J is smooth and if $t \notin (a,b) + T\mathbb{Z}$ then $J(s,t) = \gamma_{\mathbb{C}}(t)$ for all s. By proposition A.1.13 we can find h with $0 < h < \frac{b-a}{2}$ such that for all s, $J_s|_{[b-T-h,a+h]}$ is a smooth immersion with its only crossing points on $[b-T, a]$.

Find $\eta < 0$ such that $B_{2\eta}(\alpha(a))$ and $B_{2\eta}(\alpha(b))$ are disjoint. We can find $\epsilon < h$ such that $\alpha([a, a+\epsilon]) \subseteq B_\eta(\alpha(a))$, $\gamma_{\mathbb{C}}([a, a+\epsilon]) \subseteq B_\eta(\alpha(a))$ and $\alpha([b-\epsilon, b]) \subseteq B_\eta(\alpha(b))$, $\gamma_{\mathbb{C}}([b-\epsilon, b]) \subseteq B_\eta(\alpha(b))$. Thus for all s we have $J_s([a, a+\epsilon]) \subseteq B_\eta(\alpha(a))$ and $J_s([b-\epsilon, b]) \subseteq B_\eta(\alpha(b))$.

The maps $\gamma_{\mathbb{C}}|_{[a,b]}$ and $\alpha|_{[a,b]}$ are injective smooth immersions, with $\gamma_{\mathbb{C}}(a+\epsilon), \alpha(a+\epsilon) \in B_\eta(\alpha(a))$ and $\gamma_{\mathbb{C}}(b-\epsilon), \alpha(b-\epsilon) \in B_\eta(\alpha(b))$, so by proposition A.1.12 there is a smooth isotopy H from $\gamma_{\mathbb{C}}|_{[a,b]}$ to $\alpha|_{[a,b]}$ relative $(a \hookrightarrow B_\eta(\alpha(a)))$ and relative $(b \hookrightarrow B_\eta(\alpha(b)))$. Then we can find δ with $0 < \delta < \epsilon$ such that if $s \in [0,1]$ and $t \in [a+\delta, a+\epsilon]$ then $H(s,t) \in B_{2\eta}(\alpha(a))$ and if $s \in [0,1]$ and $t \in [b-\epsilon, b-\delta]$ then $H(s,t) \in B_{2\eta}(\alpha(b))$.

The next step is to "push γ into the page/pull it out of the page" on (a,b), so that in this section the knot is at a different vertical location to the rest of the knot. That means we will be able to move it around freely without hitting the rest of the knot. We can find M such that $|\gamma_3(t)| < M$ for all t. Then by proposition A.1.11 we can find a smooth knot σ smoothly isotopic to γ such that:

- $\sigma_{\mathbb{C}} = \gamma_{\mathbb{C}}$
- $\sigma(t) = \gamma(t)$ for $t \notin (a,b) + T\mathbb{Z}$
- $|\sigma_3(t)| \leqslant M$ for $t \notin (a+\delta, b-\delta)$
- $|\sigma_3(t)| \geqslant M$ for $t \in [a+\delta, b-\delta]$
- $|\sigma_3(t)| = 2M$ for $t \in [a+\epsilon, b-\epsilon]$
- $|\sigma_3|'(t) > 0$ for $t \in (a+\delta, a+\epsilon)$

- $|\sigma_3|'(t) < 0$ for $t \in (b-\epsilon, b-\delta)$

Let $g : \mathbb{R} \to \mathbb{R}$ be a smooth function with $0 \leqslant g(t) \leqslant 1$ for all t, $g(t) = 0$ for $t \leqslant a+\delta$ or $t \geqslant b-\delta$, and $g(t) = 1$ for $t \in [a+\epsilon, b-\epsilon]$. Let

$$K : [0,1] \times [b-T, a+T] \to \mathbb{R}^2,$$

$$(s,t) \mapsto \begin{cases} J(s,t) + g(t)(H(s,t) - J(s,t)) & \text{if } t \in [a,b] \\ J(s,t) & \text{if } t \notin (a+\delta, b-\delta). \end{cases}$$

This is smooth and by a similar argument to proposition A.2.8 we can extend it to a smooth function $\overline{K}$ on $[0,1] \times \mathbb{R}$ which is T-periodic in its second argument. Let $L(s,t) = \overline{K}(s,t) + \sigma_3(t)e_3$. Then L is smooth. We have $L_0 = \sigma$.

For all s and all $t \notin (a,b)+T\mathbb{Z}$ we have $L(s,t) = \gamma(t) = \sigma(t)$, and for all $t \in [a+\epsilon, b-\epsilon]$ we have $L(s,t) = H(s,t) + \sigma_3(t)$. For all s, if $t \in [a+\delta, a+\epsilon]$ then $J(s,t) \in B_\eta(\alpha(a))$ and $H(s,t) \in B_{2\eta}(\alpha(a))$ so $K(s,t) \in B_{2\eta}(\alpha(a))$. Similarly if $t \in [b-\epsilon, b-\delta]$ then for all s, $K(s,t) \in B_{2\eta}(\alpha(b))$.

Let $s \in [0,1]$. We will argue that L_s is a smooth knot. We have that $(L_s)_{\mathbb{C}}|_{[b-\delta-T, a+\delta]} = J_s|_{[b-\delta-T, a+\delta]}$, so is a smooth immersion with its only crossing points on $[a,b]$. Thus we have $L_s'(t) \neq 0$ for $t \in [b-\delta-T, a+\delta]$, and $L_s|_{(a,a+\delta]\cup[b-\delta,b)}$ is injective and if $t \in (a, a+\delta] \cup [b-\delta, b)$ then $L_s(t) \notin \text{Image}(L_s|_{[b-T,a]}$. But also $L_s|_{[b-T,a]} = \gamma|_{[b-T,a]}$ so $L_s|_{[b-T,a]}$ is injective. Thus $L_s|_{[b-\delta-T,a+\delta]}$ is injective. Then if $t \in [b-\delta-T, a+\delta]$ and $t' \in (a+\delta, b-\delta)$ we have $|\sigma_3(t)| \leqslant M$ and $|\sigma_3(t')| > M$ so $L_s(t) \neq L_s(t')$. Thus L_s has no crossing points in $[b-\delta-T, a+\delta]$.

Then if $t \in (a+\delta, a+\epsilon)$, we have $|\sigma_3(t)| \geqslant M$ and $|\sigma_3|'(t) > 0$. Thus σ_3 is injective and has non zero derivative on $(a+\delta, a+\epsilon)$, so L_s is injective and has non zero derivative on $(a+\delta, a+\epsilon)$. Then if $t \in (a+\delta, a+\epsilon)$ and $t' \in [a+\epsilon, b-\epsilon]$ we have $|\sigma_3(t)| < 2M$ and $|\sigma_3(t')| = 2M$ so $L_s(t) \neq L_s(t')$. If $t \in [a+\delta, a+\epsilon]$ and $t' \in [b-\epsilon, b-\delta]$ then $K_s(t) \in B_{2\eta}(\alpha(a))$ and $K_s(t') \in B_{2\eta}(\alpha(b))$, so by the choice of η, $K_s(t) \neq K_s(t')$. Thus

if $t \in (a+\delta, a+\epsilon)$ then t is not a crossing point of L_s. Similarly if $t \in (b-\epsilon, b-\delta)$ then t is not a crossing point of L_s. Thus L_s has no crossing points on $(b-\epsilon-T, a+\epsilon)$ and has non zero derivative on $(b-\epsilon-T, a+\epsilon)$.

Finally if $t \in [a+\epsilon, b-\epsilon]$ then $(L_s)_{\mathbb{C}}(t) = H_s(t)$, so since H_s is an injective smooth immersion we have that $(L_s)'_{\mathbb{C}}(t) \neq 0$, and if $t, t' \in [a+\epsilon, b-\epsilon]$ with $t \neq t'$ then $(L_s)_{\mathbb{C}}(t) \neq (L_s)_{\mathbb{C}}(t')$. Thus L_s has no crossing points on $[a+\epsilon, b-\epsilon]$, so has no crossing points, and on $[a-\epsilon, b+\epsilon]$ we have $L'_s(t) \neq 0$, so L_s is a smooth immersion. Thus L_s is indeed a smooth knot for all s.

Thus L is a smooth isotopy, so $L_0 = \sigma$ is isotopic to L_1, so γ is isotopic to L_1. We will take $\beta = L_1$. Then if $t \notin (a,b)+T\mathbb{Z}$ we have $\beta(t) = \gamma(t)$. Also we have $J(1,t) = \alpha(t)$ for all t, and if $t \in [a,b]$ then $H(1,t) = \alpha(t)$. Thus we have $\overline{K}_1 = \alpha$, so $\beta_{\mathbb{C}} = \alpha$, as required. □

With this in hand, to carry out the "over the shoulder" manoeuvre we just need to find a suitable target curve α. We will break this up into sections. First we find an argument θ that will (almost) play the role of Arg_α. There are quite a few conditions that this has to satisfy.

Proposition 1.15. *Let γ be a smooth knot which projects nicely, let $a < b$ such that γ has at most one backwards bend on $[a,b]$, and $D_\gamma(t) \leqslant 0$ for $t \in [a,b]$. Let $u \in (a,b]$ and suppose that $\gamma_{\mathbb{C}}$ has at most one crossing point on $[a,u]$. Let V be an open interval with $[a,u] \subseteq V$. Let $\tilde{\gamma}$ be a lifting of γ through* exp, $\mathrm{Arg}_\gamma = \tilde{\gamma}_2$.

Then there are $x, y \in V$ and a smooth function $\theta : [y-T, x] \to \mathbb{R}$ such that:

(i) $x < a$, $y > u$, *and* $y - x < \frac{T}{2}$

(ii) $\gamma_{\mathbb{C}}$ *has no crossing points on* $[x,a) \cup (u,y]$

(iii) $\theta(t) = \mathrm{Arg}_\gamma(t)$ *for* $t \leqslant x$, $\theta(t) = \mathrm{Arg}_\gamma(t) + 2\pi$ *for* $t \geqslant y$

(iv) $\theta'(t) > 0$ *for* $t \in [a,u]$

(v) $\theta'(t) > \mathrm{Arg}'_\gamma(t)$ *for* $t \in (x, y)$

(vi) $\mathrm{Arg}'_\gamma(a) = 0$ *implies* $\theta'(x) > 0$, $\mathrm{Arg}'_\gamma(a) < 0$ *implies* $\theta'(x) < 0$

(vii) $\mathrm{Arg}'_\gamma(u) = 0$ *implies* $\theta'(y) > 0$, $\mathrm{Arg}'_\gamma(u) < 0$ *implies* $\theta'(y) < 0$

(viii) θ' *has at most one zero on* $[x, a]$ *and at most one zero on* $[u, y]$

(ix) If $\mathrm{Arg}'_\gamma(u) < 0$ *then if* w *is the unique zero of* θ' *on* $[u, y]$*, if we have* $t \in [w, y]$ *and* $t' \in [y, u + T]$ *such that* $\exp(\tilde{\gamma}_1(t) + i\theta(t)) = \gamma_{\mathbb{C}}(t')$ *then we have* $t = t' = y$

(x) Any zeroes of θ' *are simple*

(xi) If $t \in [x, a]$ *then* $\theta(t) \leqslant \theta(a)$

(xii) If $t \in [u, y]$ *then* $\theta(t) \geqslant \theta(u)$

(xiii) There is $\zeta \in \mathbb{R}$ *such that* $\theta([x, y]) \subseteq (\zeta, \zeta + 2\pi)$

(xiv) The sets $\theta([x, a])$ *and* $\mathrm{Arg}_\gamma([u, b]) + 2\pi\mathbb{Z}$ *are disjoint*

Proof. We have $\mathrm{Arg}'_\gamma(t) = D_\gamma(t) \leqslant 0$ for $t \in [a, b]$, and have $\mathrm{Arg}_\gamma(u) \geqslant \mathrm{Arg}_\gamma(b) \geqslant \mathrm{Arg}_\gamma(a) - \pi$ since γ has at most one backwards bend on $[a, b]$. Thus we can find τ_1, τ_2 with $\mathrm{Arg}_\gamma(a) < \tau_1 < \tau_2 < \mathrm{Arg}_\gamma(b) + 2\pi$. Let

$$\lambda(t) = \frac{t - a}{u - a} \cdot (\tau_2 - \tau_1) + \tau_1,$$

so λ is smooth with $\lambda(a) = \tau_1 > \mathrm{Arg}_\gamma(a)$ and $\lambda(u) = \tau_2 < \mathrm{Arg}_\gamma(b) + 2\pi$. We have $\lambda'(t) = \frac{\tau_2 - \tau_1}{b - a} > 0$.

We will pick $x, y \in V$ with $x < a$ and $y > u$ to satisfy the above conditions. By picking x close enough to a and y close enough to u we can ensure that $\gamma_{\mathbb{C}}$ has no crossings on $[x, a)$ or $(u, y]$, that $y - x < \frac{T}{2}$, that $\mathrm{Arg}_\gamma(x) > \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2}$ and that $\mathrm{Arg}_\gamma(y) < \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2}$. We will add further conditions on the choice of x and y depending on whether $\mathrm{Arg}'_\gamma = 0$ at a and at u.

Suppose $\mathrm{Arg}'_\gamma(a) = 0$. Then Arg'_γ has a simple zero at a so we must have $\mathrm{Arg}''_\gamma(a) < 0$. We can choose $x < a$ so that $\mathrm{Arg}_\gamma(x) < \lambda(x)$, and such that if $t \in [x, a)$ then $\lambda'(t) > \mathrm{Arg}'_\gamma(t) > 0$. Then we have $\mathrm{Arg}_\gamma(t) < \lambda(t)$ for $t \in [x, a]$. Thus by proposition A.2.13 there is a smooth function $\theta_x : \mathbb{R} \to \mathbb{R}$ such that $\theta_x(t) = \mathrm{Arg}_\gamma(t)$ for $t \leqslant x$, $\theta_x(t) = \lambda(t)$ for $t \geqslant \frac{a+x}{2}$, $\mathrm{Arg}_\gamma(t) \leqslant \theta_x(t) \leqslant \lambda(t)$ for $t \in [x, \frac{a+x}{2}]$, and such that $\theta'_x(t) > \min(\mathrm{Arg}'_\gamma(t), \lambda'(t)) = \mathrm{Arg}'_\gamma(t)$ for $t \in [x, \frac{a+x}{2}]$. It follows that we have $\theta'_x(t) > \mathrm{Arg}'_\gamma(t)$ for $t \in (x, a]$. Also we have $\theta'_x(t) > 0$ for $t \in [x, a]$.

Now suppose $\mathrm{Arg}'_\gamma(a) < 0$. We can choose $x < a$ so that $\mathrm{Arg}_\gamma(x) < \lambda(x)$ and such that $\mathrm{Arg}'_\gamma(t) < 0$ for $t \in [x, a]$. Thus by proposition A.2.15 there is a smooth function θ_x such that $\theta_x(t) = \mathrm{Arg}_\gamma(t)$ for $t \leqslant x$, $\theta_x(t) = \lambda(t)$ for $t \geqslant a$, $\mathrm{Arg}_\gamma(t) \leqslant \theta_x(t) \leqslant \lambda(t)$ for $t \in [x, a]$, $\mathrm{Arg}'_\gamma(t) < \theta'_x(t)$ for $t \in (x, a]$, and such that θ'_x has a single simple zero in $[x, a]$.

We do a similar thing for u. Suppose $\mathrm{Arg}'_\gamma(u) = 0$. Then we cannot have $u \in \mathrm{Int}([a, b])$, and have $a > u$, so we must have $u = b$. Then since Arg'_γ has a simple zero at u we must have $\mathrm{Arg}''_\gamma(u) > 0$. We can choose $y > u$ so that if $t \in (u, y]$ then $\lambda'(t) > \mathrm{Arg}'_\gamma(t) > 0$ and such that $\mathrm{Arg}_\gamma(y) + 2\pi > \lambda(y)$. Thus by proposition A.2.13 there is a smooth function θ_y such that $\theta_y(t) = \lambda(t)$ for $t \leqslant \frac{u+y}{2}$, $\theta_y(t) = \mathrm{Arg}_\gamma(t) + 2\pi$ for $t \geqslant y$, $\lambda(t) \leqslant \theta_y(t) \leqslant \mathrm{Arg}_\gamma(t) + 2\pi$ for $t \in [\frac{u+y}{2}, y]$ and $\theta'_y(t) > \min(\mathrm{Arg}'_\gamma(t), \lambda'(t)) = \mathrm{Arg}'_\gamma(t)$ for $t \in (\frac{u+y}{2}, y)$. It follows that we have $\theta'_y(t) > \mathrm{Arg}'_\gamma(t)$ for $t \in [u, y)$, and that we have $\theta'_y(t) > 0$ for $t \in [u, y]$.

Finally suppose $\mathrm{Arg}'_\gamma(u) < 0$. Then we can choose $y > u$ so that $\mathrm{Arg}_\gamma(y) + 2\pi > \lambda(y)$ and such that $\mathrm{Arg}'_\gamma(t) < 0$ for $t \in [u, y]$. Before defining θ_y, so that we can satisfy condition (ix) we seek $v \in (u, y)$ such that if $\gamma(t) \in \exp(\tilde{\gamma}_1([v, y]) \times [\mathrm{Arg}_\gamma(y), \mathrm{Arg}_\gamma(v)])$ then $t \in [v, y] + T\mathbb{Z}$. Since y is not a crossing point of $\gamma_\mathbb{C}$, by proposition A.3.5 we can find an open interval U containing y such that $\gamma_\mathbb{C}$ has no crossing points on U. We can assume WLOG that U is small enough that if $t \in U$ then $\mathrm{Arg}'_\gamma(t) < 0$, and such that if $t, t' \in U$ then $|t - t'| < 2\pi$. Then we have $\gamma_\mathbb{C}^{-1}(\gamma_\mathbb{C}(U + T\mathbb{Z})) = U + T\mathbb{Z}$,

i.e. $U + T\mathbb{Z}$ is a saturated open set with respect to $\gamma_{\mathbb{C}}$, so by proposition A.2.6 we have that $\gamma_{\mathbb{C}}(U) = \gamma_{\mathbb{C}}(U + T\mathbb{Z})$ is open in $\gamma_{\mathbb{C}}(\mathbb{R})$. Thus we can find open intervals W, W' such that $\tilde{\gamma}(y)+2\pi i \in W \times W'$ and $\exp(W \times W') \cap \gamma_{\mathbb{C}}(\mathbb{R}) \subseteq \gamma_{\mathbb{C}}(U)$. Then we can find $v \in (u, y)$ such that $[v, y] \subseteq U$ and $\tilde{\gamma}_1([v, y]) \subseteq W$ and $[\mathrm{Arg}_\gamma(y) + 2\pi, \mathrm{Arg}_\gamma(v) + 2\pi] \subseteq W'$. Now suppose that $t \in \mathbb{R}$ is such that $\gamma_{\mathbb{C}}(t) \in \exp(\tilde{\gamma}_1([v, y]) \times [\mathrm{Arg}_\gamma(y)+2\pi, \mathrm{Arg}_\gamma(v)+2\pi])$. We seek to show that $t \in [v, y]+T\mathbb{Z}$. We have $\gamma_{\mathbb{C}}(t) \in \exp(W \times W')$ so that $\gamma_{\mathbb{C}}(t) = \gamma_{\mathbb{C}}(t')$ for some $t' \in U$, by the choice of W, W'. Then since U contains no crossing points we have $t \equiv t'$. But also $\mathrm{Arg}_\gamma(t') \in [\mathrm{Arg}_\gamma(y) + 2\pi, \mathrm{Arg}_\gamma(v) + 2\pi] + 2k\pi$ for some $k \in \mathbb{Z}$, so that $\mathrm{Arg}_\gamma(t') = \mathrm{Arg}_\gamma(t'') + 2k\pi i$ for some $t'' \in [v, y]$, and we have $t', t'' \in U$ so $|t' - t''| < 2\pi$, so $k = 0$ and $t' = t''$. Thus $t' \in [v, y]$, so indeed $t \in [v, y] + T\mathbb{Z}$.

Now let $f, g : \mathbb{R} \to \mathbb{R}$, $f : t \mapsto -\mathrm{Arg}_\gamma(-t) - 2\pi$, $g : t \mapsto -\lambda(-t)$. Then if $t \in [-y, -v]$ we have $f(t) < g(t)$, $f'(t) = \mathrm{Arg}'_\gamma(-t) < 0$ and $g'(t) = \lambda'(-t) > 0$, so we can apply proposition A.2.15 to obtain a smooth function j such that $j(t) = f(t)$ for $t \leqslant -y$, $j(t) = g(t)$ for $t \geqslant -v$, $f(t) \leqslant j(t) \leqslant g(t)$ for $t \in [-y, -v]$, $f'(t) < j'(t)$ for $t \in (-y, -v]$, and such that j' has a single simple zero on $[-y, -v]$. Letting $\theta_y : t \mapsto -j(-t)$, we have that θ_y is smooth, that $\theta_y(t) = \lambda(t)$ for $t \leqslant v$, that $\theta_y(t) = \mathrm{Arg}_\gamma(t) + 2\pi$ for $t \geqslant y$, that $\lambda(t) \leqslant \theta_y(t) \leqslant \mathrm{Arg}_\gamma(t) + 2\pi$ for $t \in [v, y]$, that $\theta'_y(t) > \mathrm{Arg}'_\gamma(t)$ for $t \in [v, y)$, and that θ'_y has a single zero in $[v, y]$. Let w be the zero of θ'_y on $[v, y]$.

We have $\theta'_y(t) \geqslant \mathrm{Arg}'_\gamma(t)$ for $t \in [w, y]$, and $\theta_y(y) = \mathrm{Arg}_\gamma(y) + 2\pi$, so we must have $\theta_y(w) \leqslant \mathrm{Arg}_\gamma(w) + 2\pi$, but also $\mathrm{Arg}_\gamma(w) \leqslant \mathrm{Arg}_\gamma(v)$ since $w \in [v, y]$ and $\mathrm{Arg}'_\gamma < 0$ on $[v, y]$. Thus $\theta_y(w) \leqslant \mathrm{Arg}_\gamma(v) + 2\pi$. Thus if we have $t \in [w, y]$ and $t' \in [y, u + T]$ with $\exp(\tilde{\gamma}_1(t) + i\theta_y(t)) = \gamma_{\mathbb{C}}(t')$ then we have $\gamma_{\mathbb{C}}(t') = \exp(\tilde{\gamma}_1(t) + i\theta_y(t)) \in \exp(\tilde{\gamma}_1([v, y]) \times [\mathrm{Arg}_\gamma(y) + 2\pi, \mathrm{Arg}_\gamma(v) + 2\pi])$ so as shown above, we must have $t' \in [v, y] + T\mathbb{Z}$. But then $t' \in [y, u + T] \cap ([v, y] + T\mathbb{Z})$, so $t' = y = t$.

Now define

$$\theta(t) = \begin{cases} \theta_x(t) & \text{if } t \in [y - T, u] \\ \theta_y(t) & \text{if } t \in [a, x + T]. \end{cases}$$

This is smooth, and satisfies $\theta(t) = \mathrm{Arg}_\gamma(t)$ for $t \leqslant x$, $\theta(t) = \mathrm{Arg}_\gamma(t) + 2\pi$ for $t \geqslant y$, and $\theta(t) = \lambda(t)$ for $t \in [a, u]$. If $t \in [a, u]$ then $\theta'(t) = \frac{\tau_2 - \tau_1}{u-a} > 0$. This means that if $t \in [a, u]$ then $\theta'(t) > \mathrm{Arg}'_\gamma(t)$, and we also have $\theta'(t) > \mathrm{Arg}'_\gamma(t)$ for $t \in (x, a]$ and $t \in [u, y)$, so if $t \in (x, y)$ then $\theta'(t) > \mathrm{Arg}'_\gamma(t)$. By the choice of x and y we have $\theta'(x) = \mathrm{Arg}'_\gamma(x) \neq 0$ and $\theta'(y) = \mathrm{Arg}'_\gamma(y) \neq 0$. By construction θ' has at most one zero on $[x, a]$ and at most one zero on $[u, y]$. It was argued above that if we have $t \in [w, y]$ and $t' \in [y, u + T]$ with $\exp(\tilde{\gamma}_1(t) + i\theta_y(t)) = \gamma(t')$ then $t = t' = y$, so that if we have $t \in [w, y]$ and $t' \in [y, u+T]$ with $\exp(\tilde{\gamma}_1(t) + i\theta(t)) = \gamma(t')$ then $t = t' = y$. That covers properties (i)–(ix) required of θ.

Any zero of θ' on $[x, a] \cup [u, y]$ is simple by construction, and θ' has no zeroes on $[a, u]$, and we have $\theta|_{[y-T,x]} = \mathrm{Arg}_\gamma|_{[y-T,x]}$ and $\theta|_{[y,x+T]} = \mathrm{Arg}_\gamma|_{[y,x+T]} + 2\pi$ so θ has only simple zeroes on $[y - T, x] \cup [y, x + T]$. Thus θ' has only simple zeroes. That covers (x). We now prove (xi)–(xiii).

First we claim that if $t \in [x, a]$ then $\min(\mathrm{Arg}_\gamma(x), \mathrm{Arg}_\gamma(a)) \leqslant \theta(t) \leqslant \max(\mathrm{Arg}_\gamma(x), \theta(a))$. Indeed suppose $\mathrm{Arg}'_\gamma(a) = 0$. Then we have $\theta'(t) > 0$ for $t \in [x, a]$, so if $t \in [x, a]$ then $\mathrm{Arg}_\gamma(x) = \theta(x) \leqslant \theta(t) \leqslant \theta(a)$, as required. Now suppose $\mathrm{Arg}'_\gamma(a) < 0$. Then we have $\mathrm{Arg}'_\gamma(t) < 0$ for $t \in [x, a]$, and if $t \in [x, a]$ we have $\theta'(t) \geqslant \mathrm{Arg}'_\gamma(t)$, so $\theta(t) - \theta(x) \geqslant \mathrm{Arg}_\gamma(t) - \mathrm{Arg}_\gamma(x)$ so $\theta(t) \geqslant \mathrm{Arg}_\gamma(t) \geqslant \mathrm{Arg}_\gamma(a)$. Moreover θ' has a single simple zero in $[x, a]$ and we have $\theta'(x) < 0$ and $\theta'(a) > 0$ so if $t \in [a, x]$ then $\theta(t) \leqslant \max(\theta(x), \theta(a)) = \max(\mathrm{Arg}_\gamma(x), \theta(a))$. That proves the claim.

Then by the choice of x we have $\mathrm{Arg}_\gamma(x) < \lambda(x)$, so $\mathrm{Arg}_\gamma(x) < \lambda(a) = \theta(a)$ so if $t \in [x, a]$ then $\theta(t) \leqslant \max(\mathrm{Arg}_\gamma(x), \theta(a)) = \theta(a)$. That proves (xi).

Next we claim that if $t \in [u, y]$ then $\min(\mathrm{Arg}_\gamma(y)+2\pi, \theta(u)) \leqslant \theta(t) \leqslant \max(\mathrm{Arg}_\gamma(y)+2\pi, \mathrm{Arg}_\gamma(u)+2\pi)$. Indeed suppose $\mathrm{Arg}'_\gamma(u) = 0$. Then we have $\theta'(t) > 0$ for $t \in [u, y]$ so if $t \in [u, y]$ then $\theta(u) \leqslant \theta(t) \leqslant \theta(y) = \mathrm{Arg}_\gamma(y)+2\pi$, as required. Now suppose $\mathrm{Arg}'_\gamma(u) < 0$. Then we have $\mathrm{Arg}'_\gamma(t) < 0$ for $t \in [u, y]$, and if $t \in [u, y]$ then $\theta'(t) \geqslant \mathrm{Arg}'_\gamma(t)$ so $\theta(y) - \theta(t) \geqslant \mathrm{Arg}_\gamma(y) - \mathrm{Arg}_\gamma(t)$ so $\theta(t) \leqslant \mathrm{Arg}_\gamma(t) + 2\pi \leqslant \mathrm{Arg}_\gamma(u) + 2\pi$. Moreover θ' has

a single simple zero in $[u, y]$, and we have $\theta'(u) > 0$ and $\theta'(y) < 0$ so if $t \in [u, y]$ then $\theta(t) \geqslant \min(\theta(y), \theta(u)) = \min(\mathrm{Arg}_\gamma(y) + 2\pi, \theta(u))$. That proves the claim.

Then by the choice of y we have $\mathrm{Arg}_\gamma(y) + 2\pi > \lambda(y)$, so $\mathrm{Arg}_\gamma(y) + 2\pi > \lambda(u) = \theta(u)$, so if $t \in [u, y]$ then $\theta(t) \geqslant \min(\mathrm{Arg}_\gamma(y) + 2\pi, \theta(u)) = \theta(u)$. That proves (xii).

It only remains to show that there is $\zeta \in \mathbb{R}$ such that $\theta([x, y]) \subseteq (\zeta, \zeta + 2\pi)$. We have that if $t \in [x, u]$ then $\theta(t) \leqslant \theta(u)$, and that if $t \in [u, y]$ then $\theta(t) \leqslant \max(\mathrm{Arg}_\gamma(y) + 2\pi, \mathrm{Arg}_\gamma(u) + 2\pi)$, so for all $t \in [x, y]$ we have $\theta(t) \leqslant \max(\mathrm{Arg}_\gamma(y) + 2\pi, \mathrm{Arg}_\gamma(u) + 2\pi)$. Similarly for all $t \in [x, y]$ we have $\theta(t) \geqslant \min(\mathrm{Arg}_\gamma(x), \mathrm{Arg}_\gamma(a))$. But by the choice of x and y we have $\mathrm{Arg}_\gamma(x) > \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2}$ and $\mathrm{Arg}_\gamma(y) < \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(b)}{2}$. Since $\mathrm{Arg}_\gamma(u) < \mathrm{Arg}_\gamma(a)$ this means that $\min(\mathrm{Arg}_\gamma(x), \mathrm{Arg}_\gamma(a)) > \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2}$. Similarly we have $\mathrm{Arg}_\gamma(u) < \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2}$, so that $\max(\mathrm{Arg}_\gamma(y) + 2\pi, \mathrm{Arg}_\gamma(u) + 2\pi) > \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2} + 2\pi$. Thus taking $\zeta = \frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2} + 2$, we have $\theta([x, y]) \subseteq (\zeta, \zeta + 2\pi)$. That proves (xiii).

We conclude with (xiv). By the above we have

$$\begin{aligned} \theta([x, a]) &\subseteq \left[\frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2}, \theta(a)\right] \\ = \left[\frac{\mathrm{Arg}_\gamma(a) + \mathrm{Arg}_\gamma(u)}{2}, \lambda(a)\right] &\subseteq (\mathrm{Arg}_\gamma(u), \mathrm{Arg}_\gamma(b) + 2\pi). \end{aligned}$$

This is disjoint of $\mathrm{Arg}_\gamma([u, b]) + 2k\pi = [\mathrm{Arg}_\gamma(b), \mathrm{Arg}_\gamma(u)] + 2k\pi$ for every $k \in \mathbb{Z}$. That proves the proposition. □

Now we find a suitable radius function for our target curve α, and combine it with the above θ.

Proposition 1.16. *Let γ be a smooth knot which projects nicely, let $a < b$ such that γ has at most one backwards bend on $[a, b]$, and $D_\gamma(t) \leqslant 0$ for $t \in [a, b]$. Let $u \in (a, b]$ and suppose that $\gamma_{\mathbb{C}}$ has at most one crossing point on $[a, u]$. Let V be an open interval with $[a, u] \subseteq V$. Let $\tilde{\gamma}$ be a lifting of γ through* exp, $\mathrm{Arg}_\gamma = \tilde{\gamma}_2$.

Then there are $x, y \in V$ and a smooth function $\phi : \mathbb{R} \to \mathbb{C}$ such that:

(i) $x < a$, $y > u$, *and* $y - x < \frac{T}{2}$

(ii) $\gamma_{\mathbb{C}}$ *has no crossing points on* $[x, a) \cup (u, y]$

(iii) $\exp \circ \phi$ *is* T*-periodic*

(iv) $(\exp \circ \phi)|_{[y-T,x]} = \gamma_{\mathbb{C}}|_{[y-T,x]}$

(v) $\phi_2'(t) > 0$ *for* $t \in [a, u]$

(vi) If $t \in (x, y)$ *then* $\phi_2'(t) > \mathrm{Arg}_\gamma'(t)$

(vii) $\phi_2'(x) \neq 0$ *and* $\phi_2'(y) \neq 0$

(viii) ϕ_2' *has at most one zero on* $[x, a]$ *and at most one zero on* $[u, y]$

(ix) Any zeroes of ϕ_2' *are simple*

(x) There is $\zeta \in \mathbb{R}$ *such that* $\phi_2([x, y]) \subseteq (\zeta, \zeta + 2\pi)$

(xi) ϕ *is an immersion*

(xii) $\phi|_{[x,y]}$ *is injective*

(xiii) If w *is the unique zero of* ϕ_2' *on* $[u, y]$, *then* $\exp \circ \phi$ *has no crossing points on* $[w, y]$

(xiv) $(\exp \circ \phi)((x, y))$ *and* $(\exp \circ \phi)([y, b])$ *are disjoint*

Proof. By proposition A.1.15 there are $x, y \in V$ and a smooth function $\theta : [y-T, x] \to \mathbb{R}$ such that:

(i) $x < a$, $y > u$, and $y - x < \frac{T}{2}$

(ii) $\gamma_{\mathbb{C}}$ has no crossing points on $[x, a) \cup (u, y]$

(iii) $\theta(t) = \mathrm{Arg}_\gamma(t)$ for $t \leqslant x$, $\theta(t) = \mathrm{Arg}_\gamma(t) + 2\pi$ for $t \geqslant y$

(iv) $\theta'(t) > 0$ for $t \in [a, u]$

(v) $\theta'(t) > \mathrm{Arg}'_\gamma(t)$ for $t \in (x, y)$

(vi) $\mathrm{Arg}'_\gamma(a) = 0$ implies $\theta'(x) > 0$, $\mathrm{Arg}'_\gamma(a) < 0$ implies $\theta'(x) < 0$

(vii) $\mathrm{Arg}'_\gamma(u) = 0$ implies $\theta'(y) > 0$, $\mathrm{Arg}'_\gamma(u) < 0$ implies $\theta'(y) < 0$

(viii) θ' has at most one zero on $[x, a]$ and at most one zero on $[u, y]$

(ix) If $\mathrm{Arg}'_\gamma(u) < 0$ then if w is the unique zero of θ' on $[u, y]$, if we have $t \in [w, y]$ and $t' \in [y, u + T]$ such that $\exp(\tilde{\gamma}_1(t) + i\theta(t)) = \gamma_{\mathbb{C}}(t')$ then we have $t = t' = y$

(x) Any zeroes of θ' are simple

(xi) If $t \in [x, a]$ then $\theta(t) \leqslant \theta(a)$

(xii) If $t \in [u, y]$ then $\theta(t) \geqslant \theta(u)$

(xiii) There is $\zeta \in \mathbb{R}$ such that $\theta([x, y]) \subseteq (\zeta, \zeta + 2\pi)$

(xiv) The sets $\theta([x, a])$ and $\mathrm{Arg}_\gamma([u, b]) + 2\pi\mathbb{Z}$ are disjoint

We will take this as the argument part of ϕ, its second co-ordinate. We only need to find a suitable function ρ to serve as the first co-ordinate of ϕ and determine the radius of $\exp \circ \phi$.

This will be formed by putting together a section ρ_x which behaves correctly around x and a section ρ_y which behaves correctly around y. We can let $r_1, r_2 \in \mathbb{R}$ with $\tilde{\gamma}_1([x, \max(y, b)]) = [r_1, r_2]$, and pick k with $k > r_2$ (one could also carry the proof through by picking $k < r_1$).

If $\mathrm{Arg}'_\gamma(a) = 0$, we can let $\rho_x : [y - T, x + T] \to \mathbb{R}$ be any smooth function with $\rho_x(t) = \tilde{\gamma}_1(t)$ for $t \leqslant x$, and $\rho_x(t) = k + 1$ for $t \geqslant a$. We have $\theta'(x) > 0$, $\theta'(a) > 0$ and θ' has at most one zero on $[a, x]$, and any zeroes are simple, so θ' can have no zeroes on

$[a, x]$, i.e. we have $\theta' > 0$ on $[a, x]$. Thus θ' is strictly increasing on $[a, x]$, so $(\rho_x + i\theta\theta)|_{[x,a]}$ is injective.

Then since $\theta' > 0$ on $[x, a]$, we have that the map $\rho + i\theta$ is injective on $[x, a]$.

If $\mathrm{Arg}'_\gamma(a) < 0$, then $\theta'(x) < 0$, $\theta'(a) > 0$ and θ' has a single simple zero on $[x, a]$. Let $w \in (x, a)$ be the point with $\theta'(w) = 0$. Let $\psi : [y - T, x + T] \to \mathbb{R}$ be a smooth function with $\psi(t) < k$ for $t \leqslant w$, $\psi(w) = k$ and $\psi'(w) > 0$, $k < \psi(t) < k + 1$ for $t \in (w, a)$ and $\psi(t) = k + 1$ for $t \geqslant a$. Then by proposition A.2.12 there is a smooth function $\rho_x : [y - T, x + T] \to \mathbb{R}$ such that $\rho_x(t) = \tilde{\gamma}_1(t)$ for $t \leqslant x$, $\rho_x(t) = \psi(t)$ for $t \geqslant w$ and $\min(\tilde{\gamma}_1(t), \psi(t)) \leqslant \rho_x(t) \leqslant \max(\tilde{\gamma}_1(t), \psi(t))$ for $t \in [x, w]$. Thus if $t \in [x, w)$ then $\tilde{\gamma}_1(t) < k$ and $\psi(t) < k$ so $\rho_x(t) < k$. We also have $\rho'_x(w) > 0$, and for $t \in [w, a]$ we have $\rho_x(t) = \psi(t) \geqslant k$, and for $t \geqslant a$ we have $\rho_x(t) = k + 1$. We claim that $(\rho_x + i\theta)|_{[a,x]}$ is injective. Indeed we have $\theta' < 0$ on $[x, w)$ and $\theta' > 0$ on $(w, a]$, so θ is strictly decreasing on $[x, w]$ and strictly increasing on $[w, a]$ and so is injective on both sets. But if $t \in [x, w)$ and $t' \in (w, a]$ then we have $\rho_x(t) < k$ and $\rho(t') \geqslant k$ by construction, so $\rho_x(t) \neq \rho_x(t')$, so $(\rho_x + i\theta)(t) \neq (\rho_x + i\theta)(t')$. Thus $(\rho_x + i\theta)|_{[x,a]}$ is indeed injective.

Similarly if $\mathrm{Arg}'_\gamma(u) = 0$, so $u = b$, we can let $\rho_y : [y - T, x + T] \to \mathbb{R}$ be any smooth function with $\rho_y(t) = \tilde{\gamma}_1(t)$ for $t \geqslant y$ and $\rho_y(t) = k + 1$ for $t \leqslant u$. Again, it is easy to see that $(\rho_y + i\theta)|_{[u,y]}$ is injective.

If $\mathrm{Arg}'_\gamma(u) < 0$ then $\theta'(y) < 0$, $\theta'(u) > 0$ and θ' has a single simple zero on $[u, y]$. Again let $w \in (u, y)$ be the point with $\theta'(w) = 0$. Let $\eta : [x - T, y + T] \to \mathbb{R}, t \mapsto \tilde{\gamma}_1(t) + i\theta(t)$. By (ix) we have that if $t \in [w, y]$, $t \neq y$ then $\exp(\eta(t)) \notin \gamma_{\mathbb{C}}([y, u + T])$. Pick $v \in (w, y)$. Then we have for $t \in [w, v]$ that $t \neq y$ so $\exp(\eta(t)) \notin \gamma_{\mathbb{C}}([y, u+T])$, and so there is $\epsilon > 0$ such that $|\exp(\eta(t)) - \gamma_{\mathbb{C}}(s)| \geqslant \epsilon$ for all $t \in [w, v]$, $s \in [y, u+T]$. Then we can find $\delta \in (0, 1)$ such that if $|x - \eta(t)| < \delta$ for some $t \in [w, v]$ then $|\exp(x) - \exp(\eta(t))| < \epsilon$, so $\exp(x) \notin \gamma_{\mathbb{C}}([y, u + T])$. Next we can find $v' \in (w, v)$ such that $\tilde{\gamma}_1([w, v']) \subseteq (c, c + \frac{\delta}{2})$ for some $c \in \mathbb{R}$. Necessarily we have $c \leqslant k$. Finally we can find $z \in (u, w)$ such that $\theta(z) > \theta(v')$.

Then we can find a smooth function $\chi : [y-T, x+T] \to \mathbb{R}$ such that $\chi(t) = k+1$ for $t \leqslant z$, such that $\chi(t) \in (c+\frac{\delta}{2}, k+1)$ if $t \in (z,w)$, such that $\chi(w) = c+\frac{\delta}{2}$ and $\chi'(w) < 0$, and such that if $t \in (w, v']$ then $\chi(t) \in (c, c+\frac{\delta}{2})$. Then by proposition A.2.12 there is a smooth function $\rho_y : [y-T, x+T] \to \mathbb{R}$ such that $\rho_y(t) = \chi(t)$ for $t \leqslant w$, $\rho_y(t) = \tilde{\gamma}_1(t)$ for $t \geqslant v'$, and $\min(\tilde{\gamma}_1(t), \chi(t)) \leqslant \rho_y(t) \leqslant \max(\tilde{\gamma}_1(t), \chi(t))$ for $t \in [w, v']$. Thus if $t \in (w, v']$ then $\rho_y(t) \in (c, c+\frac{\delta}{2})$, and we have $\rho_y'(w) < 0$, and $\rho_y(t) \geqslant c+\frac{\delta}{2}$ for $t \in [z, w]$, and $\rho_y(t) = k+1$ for $t \leqslant z$.

Again we claim that $(\rho_y + i\theta)|_{[u,y]}$ is injective. Indeed, we have that θ is strictly increasing on $[u, w]$ and strictly decreasing on $[w, y]$, so is injective on both sets, so $(\rho_y + i\theta)$ is injective on $[u, w]$ and on $[w, y]$. Suppose for contradiction that $t \in [u, w]$ and $t' \in (w, y]$ with $\rho_y(t) + i\theta(t) = \rho_y(t') + i\theta(t')$. We have $\rho_y(t) = \chi(t) \geqslant c+\frac{\delta}{2}$, so if we had $t' \in (w, v']$ then $\rho_y(t) \neq \rho_y(t')$, so we must have $t' \in (v', y]$. But then $\theta(t') \leqslant \theta(v') < \theta(z)$ by the choice of z, and if $t \in [z, w]$ then we have $\theta(t) \geqslant \theta(z)$, so if $t \in [z, w]$ then we would have $\theta(t) > \theta(t')$. Thus we cannot have $t \in [z, w]$, so we must have $t \in [u, z)$. But then $\rho_y(t) = k+1 > \rho_y(t')$, a contradiction. Thus no such t and t' exist, so $(\rho_y + i\theta)|_{[u,y]}$ is injective as claimed.

Now define

$$\rho : [y-T, x+T] \to \mathbb{R},\ t \mapsto \begin{cases} \rho_x(t) & \text{if } t \leqslant u \\ \\ \rho_y(t) & \text{if } t \geqslant a. \end{cases}$$

This is well defined and smooth by construction. Thus $\zeta : [y-T, x+T] \to \mathbb{C},\ t \mapsto \rho(t) + i\theta(t)$ is a smooth map, and $\exp \circ \zeta$ is periodic on $[y-T, x+T]$. Thus by proposition A.3.22 we obtain that $\exp \circ \zeta$ extends to a T-periodic curve τ. Then τ has a lifting ϕ through exp which satisfies $\phi|_{[y-T,x+T]} = \zeta = \rho + i\theta$. Thus $\phi : \mathbb{R} \to \mathbb{C}$ is smooth with $\phi|_{[y-T,x+T]} = \rho + i\theta$, and $\exp \circ \phi$ is T-periodic. By construction we have $\phi|_{[y-T,x]} = \tilde{\gamma}|_{[y-T,x]}$, so $(\exp \circ \phi)|_{[y-T,x]} = \gamma_{\mathbb{C}}$. Also, since $\phi_2|_{[y-T,x+T]} = \theta$ we have that if $t \in (x, y)$ then $\phi_2'(t) > \mathrm{Arg}_\gamma'(t)$, that $\phi_2'(x) \neq 0$ and $\phi_2'(y) \neq 0$, that ϕ_2' has at most one zero on $[x, a]$ and at most one zero on $[u, y]$, and that any zeroes of ϕ_2' are simple. Thus this choice of

ϕ, x and y satisfies properties (i)–(x) of those required by the proposition.

Next we show that ϕ is a smooth immersion. Indeed as just noted, $(\exp\circ\phi)|_{[y-T,x]} = \gamma_{\mathbb{C}}$ is a smooth immersion, so $\phi|_{[y-T,x]}$ must be a smooth immersion since $(\exp\circ\phi)'(t) = \phi'(t)\exp(\phi(t))$. But also if $t \in [x,y]$ with $\theta'(t) = 0$ then t is either the unique zero of θ' in $[x,a]$ or the unique zero of θ' in $[u,y]$, and either way we have by the construction of ρ that $\rho'(t) \neq 0$. Thus ϕ' has no zeroes on $[x,y]$, so has no zeroes on $[y-T,y]$, so has no zeroes. That proves (xi).

Now for (xii). We have $\theta'(t) > 0$ for $t \in (a,u)$, so if $t \in (a,u)$ then $\theta(a) < \theta(t) < \theta(u)$. But by (xi) and (xii) for θ, if $t \in [x,a]$ then $\theta(t) \leqslant \theta(a)$, and if $t \in [u,y]$ then $\theta(t) \geqslant \theta(u)$. Thus $\theta([x,a])$, $\theta((a,u))$ and $\theta([u,y])$ are disjoint, so $\phi([x,a])$, $\phi((a,u))$ and $\phi([u,y])$ are disjoint. Thus to show that $\phi|_{[x,y]}$ is injective, it suffices to show that $\phi|_{[x,a]}$, $\phi|_{(a,u)}$ and $\phi|_{[u,y]}$ are injective. But the middle of these is obvious, and the first and third have already been shown. That proves (xii).

Now for (xiii). Suppose that ϕ_2' has a zero on $[u,y]$, necessarily unique. Then $\mathrm{Arg}'_\gamma(u) < 0$. Let w, v, ϵ, δ, and v' be as described above in the construction of ρ_y in this case. As shown above we have that $\phi|_{[x,y]}$ is injective, so $(\exp\circ\phi)|_{[x,y]}$ is injective since there is $\zeta \in \mathbb{R}$ such that $\phi_2([x,y]) \subseteq (\zeta, \zeta + 2\pi)$. Thus if $t \in [w,y]$ was a crossing point of $\exp\circ\phi$ there would be $t' \in (y, x+T)$ with $\exp(\phi(t)) = \exp(\phi(t')) = \gamma_{\mathbb{C}}(t')$. But if $t \in [v',y]$ then $\phi(t) = \tilde{\gamma}_1(t) + i\theta(t)$, and so by (ix) if $\exp(\phi(t)) = \gamma_{\mathbb{C}}(t')$ then $t = t' = y$ so (t,t') is not a crossing pair of $\exp\circ\phi$. And if $t \in [w,v']$ then we have

$$|\phi(t) - \eta(t)| = |(\rho_y(t) + i\theta(t)) - (\tilde{\gamma}_1(t) + i\theta(t)| = |\rho_y(t) - \tilde{\gamma}_1(t)| \leqslant \frac{\delta}{2} < \delta$$

since we have both $\rho_y(t) \in [c, c+\frac{\delta}{2}]$ and $\tilde{\gamma}_1(t) \in [c, c+\frac{\delta}{2}]$. Thus by the choice of δ we have that $|\exp(\phi(t)) - \exp(\eta(t))| < \epsilon$, so by the choice of ϵ we have $\exp(\phi(t)) \notin \gamma_{\mathbb{C}}([y, u+T])$. Thus there is no crossing pair (t,t') with $t \in [w,v']$ and $t' \in [y, x+T]$. Thus $\exp\circ\phi$ has no crossing points on $[w,y]$, proving (xiii).

Finally, we show (xiv). If $u = b$ then $y > b$ so the conclusion is trivial, so we may assume $u < b$. Thus $\mathrm{Arg}'_\gamma(u) < 0$. We have $(\exp\circ\phi)([y,b]) = \gamma([y,b])$. Then by (xiv) for θ, we have that $\phi_2((x,a])$ is disjoint from $\mathrm{Arg}_\gamma([y,b]) + 2\pi\mathbb{Z}$, so $(\exp\circ\phi)((x,a])$ is disjoint from $\gamma([y,b])$. Letting z be as described above in the construction of ρ_y, if $t \in [a,z]$ then $\phi_1(t) = \rho(t) = k+1$, which is not in $\tilde{\gamma}_1([y,b])$ by the choice of k, so $(\exp\circ\phi)([a,z])$ is disjoint from $\gamma([y,b])$. Finally, we cover $(\exp\circ\phi)([z,y))$. If we had $t \in [w,y)$ with $\exp(\phi(t)) \in (\exp\circ\phi)([y,b])$ then t would be a crossing point of $\exp\circ\phi$, contradicting (xiii) which was just proved. Thus we need only cover $t \in (z,w)$. But if $t \in [w,z]$ then $\mathrm{Arg}_\gamma(w) \geqslant \theta(w) \geqslant \theta(t) \geqslant \theta(z) > \theta(v') > \theta(y)$ so again $\exp(\phi(t)) \notin \gamma([y,b])$. Thus indeed, $(\exp\circ\phi)((x,y))$ is disjoint from $(\exp\circ\phi)([y,b])$.

That proves the proposition. □

This $\exp\circ\phi$ is almost suitable as the plane projection of a smooth knot which has the section $[a,u]$ of γ bent forwards without adding crossings to $[a,b]$. The only problem is that $\exp\circ\phi$ may not be regular. We can fix this using proposition A.3.23.

Proposition 1.17. *Let γ be a smooth knot which projects nicely, let $a < b$ such that γ has at most one backwards bend on $[a,b]$, and $D_\gamma(t) \leqslant 0$ for $t \in [a,b]$. Let $u \in (a,b]$ and suppose that $\gamma_\mathbb{C}$ has at most one crossing point on $[a,u]$. Let V be an open interval with $[a,u] \subseteq V$. Let $\tilde{\gamma}$ be a lifting of γ through* exp, *$\mathrm{Arg}_\gamma = \tilde{\gamma}_2$.*

Then there are $x, y \in V$ and a smooth function $\tilde{\alpha} : \mathbb{R} \to \mathbb{C}$ such that, letting $\alpha = \exp\circ\tilde{\alpha}$, we have:

(i) $x < a$, $y > u$, *and* $y - x < \frac{T}{2}$

(ii) $\gamma_\mathbb{C}$ *has no crossing points on* $[x,a) \cup (b,y]$

(iii) $\alpha|_{[y-T,x]} = \gamma_\mathbb{C}|_{[y-T,x]}$

(iv) $\mathrm{Arg}'_\alpha(t) > 0$ *for* $t \in [a,u]$

(v) If $t \in (x,y)$ *then* $\mathrm{Arg}'_\alpha(t) > \mathrm{Arg}'_\gamma(t)$

(vi) α *is regular*

(vii) Arg'_α *has at most one zero on* $[x,a]$ *and at most one zero on* $[u,y]$

(viii) $\mathrm{Arg}'_\alpha(t)$ *has only simple zeroes*

(ix) $\alpha|_{[x,y]}$ *is injective*

(x) If w *is the unique zero of* Arg'_α *on* $[u,y]$*, then* α *has no crossing points on* $[w,y]$

(xi) $\alpha((x,y))$ *and* $\alpha([y,b])$ *are disjoint*

Proof. By proposition A.1.16 there are $x, y \in V$ and a smooth function $\phi : \mathbb{R} \to \mathbb{C}$ such that:

(i) $x < a$, $y > u$, and $y - x < \frac{T}{2}$

(ii) $\gamma_{\mathbb{C}}$ has no crossing points on $[x,a) \cup (u,y]$

(iii) $\exp \circ \phi$ is T-periodic

(iv) $(\exp \circ \phi)|_{[y-T,x]} = \gamma_{\mathbb{C}}|_{[y-T,x]}$

(v) $\phi_2'(t) > 0$ for $t \in [a,u]$

(vi) If $t \in (x,y)$ then $\phi_2'(t) > \mathrm{Arg}'_\gamma(t)$

(vii) $\phi_2'(x) \neq 0$ and $\phi_2'(y) \neq 0$

(viii) ϕ_2' has at most one zero on $[x,a]$ and at most one zero on $[u,y]$

(ix) Any zeroes of ϕ_2' are simple

(x) There is $\zeta \in \mathbb{R}$ such that $\phi_2([x,y]) \subseteq (\zeta, \zeta + 2\pi)$

(xi) ϕ is an immersion

(xii) $\phi|_{[x,y]}$ is injective

(xiii) If w is the unique zero of ϕ_2' on $[u, y]$, then $\exp \circ \phi$ has no crossing points on $[w, y]$

(xiv) $(\exp \circ \phi)((x, y))$ and $(\exp \circ \phi)([y, b])$ are disjoint

We will use proposition A.3.23 to perturb a section of ϕ so that $\exp \circ \phi$ is regular. We let $\tau = \exp \circ \phi$. First note that since $\phi|_{[x,y]}$ is injective, and there is $\zeta \in \mathbb{R}$ such that $\phi_2([x, y]) \subseteq (\zeta, \zeta + 2\pi)$, we have that $\tau|_{[x,y]}$ is injective. Next note that since ϕ is an immersion, so is τ, so by proposition A.3.5 the set of crossing points of τ is a closed set.

Now we claim that x is not a crossing point of τ. Indeed if $t \in [x, y]$ then since $\tau|_{[x,y]}$ is injective, (x, t) is not a crossing pair of τ. If $t \in [y - T, x)$ then $\tau(x) = \gamma_{\mathbb{C}}(x)$ and $\tau(t) = \gamma_{\mathbb{C}}(t)$, and x is not a crossing point of $\gamma_{\mathbb{C}}$, so $\gamma_{\mathbb{C}}(x) \neq \gamma_{\mathbb{C}}(t)$ and (x, t) is not a crossing pair of τ. Thus indeed, x is not a crossing point of τ. Similarly y is not a crossing point of τ. Thus since crossing points of τ form a closed set we can find $h > 0$ such that τ has no crossing points on $[x, x + h] \cup [y - h, y]$. WLOG we may assume that $x + h < a$ and $y - h > u$, and that $\mathrm{Sign}(\phi_2'(x + h)) = \mathrm{Sign}(\phi_2'(x))$, and $\mathrm{Sign}(\phi_2'(y - h)) = \mathrm{Sign}(\phi_2'(y))$.

Now pick $x' \in (x, x + h)$, $y' \in (y - h, y)$. We have $\zeta \in \mathbb{R}$ such that $\phi_2([x, y]) \subseteq (\zeta, \zeta + 2\pi)$, and can find $\epsilon_1 > 0$ such that $\phi_2([x, y]) \subseteq (\zeta + \epsilon_1, \zeta + 2\pi - \epsilon_1)$. Since $\phi_2'(t) > 0$ for $t \in [a, u]$ we can find $\epsilon_2 > 0$ such that $\phi_2'(t) > \epsilon_2$ for $t \in [a, u]$. Then since $\phi_2'(t) > \mathrm{Arg}_\gamma'(t)$ for $t \in [x', y']$ we can find $\epsilon_3 > 0$ such that $\phi_2'(t) > \mathrm{Arg}_\gamma'(t) + \epsilon_3$ for $t \in [x', y']$. Next, we have $\mathrm{Sign}(\phi_2(x + h)) = \mathrm{Sign}(\phi_2(x))$, so since any zeroes of ϕ_2' are simple and ϕ_2' has at most one zero on $[x, a]$ we have that $|\phi_2'|(t) > 0$ for $t \in [x, x + h]$. Similarly we have $|\phi_2'|(t) > 0$ for $t \in [y - h, y]$. Thus there is $\epsilon_4 > 0$ such that $|\phi_2'|(t) > \epsilon_4$ for $t \in [x, x + h] \cup [y - h, y]$. After that since $\tau(x') \notin \tau([y - T, x])$ and $\tau(y') \notin \tau([y - T, x])$, there is $\eta > 0$ such that $B_\eta(\tau(x')) \cup B_\eta(\tau(y'))$ is disjoint from $\tau([y - T, x])$. Then we can find ϵ_5 such that if $|z - \phi(x')| < 2\epsilon_5$ then $\exp(z) \in B_\eta(\tau(x'))$ and if $|z - \phi(y')| < 2\epsilon_5$ than $\exp(z) \in B_\eta(\tau(u'))$.

Next suppose ϕ_2' has a zero w on $[u, y]$. This is necessarily unique, and we have

$w \in (u, y)$. By property (xiii), we have that $\tau([w, y'])$ and $\tau([y - T, x])$ are disjoint, so since these are compact we can find $\eta_2 > 0$ such that if $s \in [w, y']$ and $t \in [y - T, x]$ then $|\tau(s) - \tau(t)| > \eta_2$. Then we can find $\epsilon_6 > 0$ such that if $s \in [w, y']$ and $|z - \phi(s)| < \epsilon_6$ then $|\exp(z) - \tau s| < \eta_2$, so $\exp(z) \notin \tau([y - T, x])$. If ϕ_2' does not have a zero on $[u, y]$, set ϵ_6 to be any positive value.

Finally we have that $\tau((x, y))$ and $\tau([y, b])$ are disjoint, so $\tau([x', y']$ and $\tau([y, b])$ are disjoint, so similarly we can find $\epsilon_7 > 0$ such that if $s \in [x', y']$ and $|z - \phi(s)| < \epsilon_7$ then $\exp(z) \notin \tau([y, b])$. Let $\epsilon = \min(\epsilon_1, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_6, \epsilon_7) > 0$.

Let $\delta > 0$ such that if $t \in [x', x' + \delta]$ then $\phi(t) \in B_\epsilon(\phi(x'))$, and if $t \in [y' - \delta, y']$ then $\phi(t) \in B_\epsilon(\phi(y'))$. We also require that $x' + \delta < x + h$ and $y' - \delta > y - h$. Thus we have $\delta < \frac{y' - x'}{2}$.

For each $j \in \mathbb{Z}$, let $\sigma_j : [y - T, x] \to \mathbb{R}^2$, $t \mapsto \tilde{\gamma}(t) + 2\pi j i$. Let X be the set of points $\tilde{\gamma}(t) + 2\pi j i \in \mathbb{R}$ such that t is a crossing point of γ and $j \in \mathbb{Z}$, which is a countable set. Then by proposition A.3.23 applied to the injective smooth immersion $\phi|_{[x,y]}$, the countable family $(\sigma_j)_{j \in \mathbb{Z}}$, the countable set X, the points $x' < y' \in [x, y]$, and this choice of δ and ϵ, there is an injective smooth immersion $\beta : [x, y] \to \mathbb{R}^2$ with $\beta(t) = \phi(t)$ for $t \notin (x', y')$, $\|\beta - \phi\|_{C^1} < \epsilon$, $\beta'(t) = \phi'(t)$ for $t \in [x' + \delta, y' - \delta]$, and such that for all $t \in [x' + \delta, y' - \delta]$ we have $\beta(t) \notin X$, and if $\beta(t) = \sigma_j(s)$ for any j then $\beta'(t)$ and $\sigma_j'(s)$ are not parallel.

Now if $t \in [x, y]$ then $|\beta_2(t) - \phi_2(t)| \leqslant \|\beta - \phi\|_{C^1} < \epsilon < \epsilon_1$ so since $\phi_2(t) \in (\zeta + \epsilon_1, \zeta + 2\pi - \epsilon_1)$ we have $\beta_2(t) \in (\zeta, \zeta + 2\pi)$. Thus $\beta_2([x, y]) \subseteq (\zeta, \zeta + 2\pi)$, so since β is injective we have that $\exp \circ \beta$ is injective.

Next if $t \in [a, u]$ then we have $|\beta_2'(t) - \phi_2'(t)| \leqslant \|\beta - \phi\|_{C^1}$ so $\beta_2'(t) > \phi_2'(t) - \epsilon_2 > 0$, by the choice of ϵ_2.

Next if $t \in [x', y']$ then similarly we obtain $\beta_2'(t) > \phi_2'(t) - \epsilon_3 > \mathrm{Arg}_\gamma'(t)$ by the choice of ϵ_3. Also if $x \in (x, x'] \cup [y', y)$ then $\beta_2'(t) = \phi_2'(t) > \mathrm{Arg}_\gamma'(t)$. Thus for all $t \in (x, y)$ we have $\beta_2'(t) > \mathrm{Arg}_\gamma'(t)$.

Next if $t \in [x, x+h] \cup [y-h, y]$ then we have $|\beta_2'(t) - \phi_2'(t)| < \epsilon_4$ and $|\phi_2'|(t) > \epsilon_4$ so $\beta_2'(t) \neq 0$.

Then if $t \in [x', x'+\delta]$ then $\|\beta(t) - \phi(x')\| \leqslant \|\beta - \phi\|_{C^1} + \|\phi(t) - \phi(x')\| < 2\epsilon \leqslant 2\epsilon_5$, so $\exp(\beta(t)) \in B_\eta(\tau(x'))$, so so $\exp(\beta(t)) \notin \tau([y-T, x])$. Thus $\exp(\beta([x', x'+\delta]))$ is disjoint from $\tau([y-T, x])$. But also if $s \in (x, x']$ and $t \in [y-T, x]$ then $\tau(s) \neq \tau(t)$, since τ has no crossing points on $[x, x+h]$, so $\exp(\beta(s)) = \exp(\phi(s)) \neq \tau(t)$. Thus $\exp(\beta((x, x'+\delta]))$ and $\tau([y-T, x])$ are disjoint. Similarly $\exp(\beta([y'-\delta, y))$ and $\tau([y-T, x])$ are disjoint.

Next suppose ϕ_2' has a zero w on $[u, y]$. Then if $t \in [w, y']$ we have $\|\beta(t) - \phi(t)\| < \epsilon_6$ so by the choice of ϵ_6, $\exp(\beta(t)) \notin \tau([y-T, x])$. Thus by the above, $\exp(\beta[w, y))$ and $\tau([y-T, x])$ are disjoint.

Finally note that if $t \in [x', y']$ then $\|\beta(t) - \phi(t)\| < \epsilon_7$ so by the choice of ϵ_7, $\exp(\beta(t)) \notin \tau([y, b])$. But also if $t \in (x, x'] \cup [y', y)$ then $\exp(\beta(t)) = \exp(\phi(t)) = \tau(t) \notin \tau([y, b])$. Thus we have that $\exp(\beta((x, y))$ and $\tau([y, b])$ are disjoint.

Since $\beta(t) = \phi(t)$ for $t \in [x, x'] \cup [y', y]$, the map

$$\zeta : [y-T, x+T] \to \mathbb{R},\ t \mapsto \begin{cases} \beta(t) & \text{if } t \in [x, y] \\ \phi(t) & \text{if } t \in [y-T, x'] \cup [y', x+T] \end{cases}$$

is smooth. This is a smooth immersion since β and ϕ both are. Then $\exp \circ \zeta$ is T-periodic, so has an extension α to $\mathbb{R}$. Let $\tilde{\alpha}$ be a lifting of α through exp, with $\tilde{\alpha}|_{[x-T, y+T]} = \zeta$. We have $\alpha|_{[y-T, x]} = \exp \circ \phi|_{[y-T, x]} = \tau|_{[y-T, x]} = \gamma|_{[y-T, x]}$.

Note that $\alpha|_{[x, y]} = \exp \circ \beta$ is injective. Thus α has no crossing pairs (s, t) with $s, t \in [x, y]$. Also if $s \in (x, x'+\delta]$ and $t \in [x-T, y]$ then $\alpha(s) = \exp(\beta(s)) \neq \tau(t) = \alpha(t)$ since $\exp(\beta((x, x'+\delta]))$ and $\tau([y-T, x])$ are disjoint. Thus α has no crossing points on $(x, x' + +\delta]$. Also taking $s = x$ and $t \in [y-T, x]$ we have $\alpha(x) = \tau(s)$ and $\alpha(t) = \tau(t)$, and x is not a crossing point of τ, so (x, t) is not a crossing pair of α. Thus α has no crossing points on $[x, x'+\delta]$. Similarly it has no crossing points on $[y'-\delta, y]$.

We claim that this map $\tilde{\alpha}$ satisfies the requirements of the proposition. We have that $\alpha = \exp \circ \tilde{\alpha}$, and obviously have properties (i) and (ii) satisfied. We have $\alpha|_{[x-T,y]} = \exp \circ \zeta|_{[x-T,y]} = \exp \circ \phi_{[x-T,y]} = \gamma_{\mathbb{C}}|_{[x-T,y]}$, giving property (iii).

Showing (iv) is straightforward, since if $t \in [a, uy]$ then $\mathrm{Arg}'_{\alpha}(t) = \beta_2'(t) > 0$, as argued above. Then (v) is also straightforward, since if $t \in (x, y)$ then $\mathrm{Arg}'_{\alpha}(t) = \beta_2'(t) > \mathrm{Arg}'_{\gamma}(t)$ as argued above.

Now for (vi). We have that $\tilde{\alpha}|_{[x-T,y+T]} = \zeta$, which is a smooth immersion, so $\tilde{\alpha}$ is a smooth immersion, so α is a smooth immersion. If (s, t) is a crossing pair of α then as noted above we cannot have both $s \in [x, y]$ and $t \in [x, y]$. Thus if α had any triple crossing (r, s, t), we could not have two of $\{r, s, t\}$ in $[x, y]$, so WLOG can assume that $r, s \in [y - T, x]$, $r \neq s$. Since $\alpha|_{[x-T,y]} = \gamma$ we have that (r, s) is a crossing point of $\gamma_{\mathbb{C}}$, and we cannot also have $t \in [y - T, x] + T\mathbb{Z}$ (or $\gamma_{\mathbb{C}}$ would have a triple crossing), so we can assume $t \in [x, y]$. But as noted above, α has no crossing points on $[x, x' + \delta] \cup [y' - \delta, y]$, so in fact we may assume $t \in (x' + \delta, y' - \delta)$. Then we have $\alpha(t) = \exp(\beta(t)) = \gamma_{\mathbb{C}}(r) = \exp \tilde{\gamma}(r)$, so there is $j \in \mathbb{Z}$ such that $\beta(t) = \tilde{\gamma}(t) + 2\pi j i$, so $\beta(t) \in X$. But that contradicts the construction of β. Thus α has no triple crossings.

Now to show that any crossing pairs of α are transversal. If (s, t) is a crossing pair of α with $s, t \in [y - T, x]$, then we have $\alpha'(s) = \gamma'_{\mathbb{C}}(s)$ and $\alpha'(t) = \gamma'_{\mathbb{C}}(t)$ so (s, t) is transversal. If $s \in [x, y]$ and $t \in [x, y]$ then (s, t) is not a crossing pair of of α. Finally suppose (s, t) is a crossing pair of α with $s \in [x, y]$ and $t \in [y - T, x]$. Thus necessarily $s \in (x' + \delta, y' - \delta)$. Then we have $\exp(\beta(s)) = \alpha(s) = \gamma(t)$, so $\beta(s) = \tilde{\gamma}(t) + 2\pi j i$ for some $j \in \mathbb{Z}$, i.e. $\beta(s) = \sigma_j(t)$. Thus by the construction of β we have that $\beta'(s)$ and $\sigma'_j(t)$ are not parallel. Thus $\beta'(s)$ and $\tilde{\gamma}'(t)$ are not parallel, so $(\exp \circ \beta)'(s) = \beta'(s) \exp(\beta(s))$ is not parallel to $(\exp \circ \tilde{\gamma})'(t) = \tilde{\gamma}'(t)\gamma(t)$, since $\exp(\beta(s)) = \gamma(t)$. Thus crossing pairs of α are indeed transversal. Thus α is indeed regular, proving (vi).

Now for (vii). We have $\mathrm{Arg}_{\alpha}|_{[x,y]} = \beta_2$. As noted above, if $t \in [x, x + h] \cup [y - h, y]$ then $\beta_2'(t) \neq 0$. We also have that if $t \in [x' + \delta, y' - \delta]$ then $\beta_2(t) = \phi_2(t)$, so that since

$x' + \delta < x + h$ and $y' - \delta > y - h$, any zero of β_2' on $[x, y]$ is also a zero of ϕ_2', and lies in either $(x + h, a)$ or in $(u, y - h)$. Since ϕ_2' has at most one zero on $[x, a]$ and at most one zero on $[u, y]$, this proves (vii).

For (viii), we have $\alpha|_{[y-T,x]} = \gamma$, so any zero of Arg'_α on $[y - T, x]$ is a simple zero. Thus we need only check that zeroes of $\mathrm{Arg}'_\alpha|_{[x,y]} = \beta_2'$ are simple. But if $w \in [x, y]$ with $\beta_2'(w) = 0$ then as just argued we have $w \in (x + h, a) \cup (u, y - h)$, and we have that w is also a zero of ϕ_2', so $\phi_2''(w) \neq 0$. But also we have by the construction of β that $\beta'|_{[x'+\delta,y'-\delta]} = \phi'_{[x'+\delta,y'-\delta]}$, so that $\beta_2''(w) = \phi_2''(w) \neq 0$, as required.

For (ix), we have $\alpha|_{[x,y]} = \exp \circ \beta$ which is injective as noted above.

For (x), suppose that $\mathrm{Arg}_\prime$ has a zero $w \in [u, y]$, necessarily unique by (vii). Then $\alpha([w, y)) = \exp(\beta([w, y))$ and $\alpha([y-T, x]) = \tau([y-T, x])$ are disjoint, as noted previously, so if $s \in [w, y)$ and $t \in [y - T, x]$ then (s, t) is not a crossing pair of α. Also $\alpha|_{[x,y]}$ is injective, so if $s \in [w, y)$ and $t \in [x, y]$ then (s, t) is not a crossing pair of α. Thus if $s \in [w, y)$ then s is not a crossing point of α. But it was noted above that y is not a crossing point of α. Thus indeed, α has no crossing points on $[w, y]$, proving (x).

Finally, (xi) is easy: we have $\alpha((x, y)) = \exp(\beta((x, y))$ which is disjoint from $\tau([y, b]) = \alpha([y, b])$ as noted above. □

Now the hard work is done. To obtain the Bending Forward proposition we need one final definition.

Definition 1.18. Let γ be a smooth knot which projects nicely. Let $c \leqslant a \leqslant b \leqslant d$. Let I be a compact interval. Say that a smooth knot β is a **bending forwards of $(\gamma, [a, b])$ on I without adding crossings to $[c, d]$** if β is a bending forwards of γ on I such that:

1. $D_\beta(t) > 0$ for $t \in [a, b]$
2. If $s \in [c, d]$ has $D_\beta(s) \leqslant 0$ and s is a crossing point of $\beta_{\mathbb{C}}$ then s is a crossing point of $\gamma_{\mathbb{C}}$

Now we can put propositions A.1.14 and A.1.17 to use.

Proposition 1.19 (Bending forwards one crossing point)**.** *Let γ be a smooth knot which projects nicely, let $a < b$ such γ has at most one backwards bend on $[a, b]$, and such that $D_\gamma(t) \leqslant 0$ for $t \in [a, b]$. Let $u \in (a, b]$ and suppose that $\gamma_{\mathbb{C}}$ has at most one crossing point on $[a, u]$. Suppose that V is an open interval containing $[a, u]$. Then there is a compact interval I with $[a, u] \subseteq I \subseteq V$ and a smooth knot β such that β projects nicely and is a bending forwards of $(\gamma, [a, u])$ on I without adding crossings to $[a, b]$.*

Proof. Let γ be a smooth knot which projects nicely, let $a < b$ such that γ has at most one backwards bend on $[a, b]$, and $D_\gamma(t) \leqslant 0$ for $t \in [a, b]$. Let $u \in (a, b]$ and suppose that $\gamma_{\mathbb{C}}$ has at most one crossing point on $[a, u]$. Let V be an open interval with $[a, u] \subseteq V$. Let $\tilde{\gamma}$ be a lifting of γ through exp, $\mathrm{Arg}_\gamma = \tilde{\gamma}_2$.

Then by proposition A.1.17 there are $x, y \in V$ and a smooth function $\tilde{\alpha} : \mathbb{R} \to \mathbb{C}$ such that, letting $\alpha = \exp \circ \tilde{\alpha}$, we have:

(i) $x < a$, $y > u$, and $y - x < \frac{T}{2}$

(ii) $\gamma_{\mathbb{C}}$ has no crossing points on $[x, a) \cup (b, y]$

(iii) $\alpha|_{[y-T,x]} = \gamma_{\mathbb{C}}|_{[y-T,x]}$

(iv) $\mathrm{Arg}'_\alpha(t) > 0$ for $t \in [a, u]$

(v) If $t \in (x, y)$ then $\mathrm{Arg}'_\alpha(t) > \mathrm{Arg}'_\gamma(t)$

(vi) $\alpha \in C^\infty_T(\mathbb{R}, \mathbb{C})$ is regular

(vii) Arg'_α has at most one zero on $[x, a]$ and at most one zero on $[u, y]$

(viii) $\mathrm{Arg}'_\alpha(t)$ has only simple zeroes

(ix) $\alpha|_{[x,y]}$ is injective

(x) If w is the unique zero of Arg'_α on $[u, y]$, then α has no crossing points on $[w, y]$

(xi) $\alpha((x,y))$ and $\alpha([y,b])$ are disjoint

Then since $\gamma_{\mathbb{C}}$ has at most one crossing point on (x,y), and $\alpha|_{[x,y]}$ is injective, by proposition A.1.14 there is a smooth knot β which is smoothly isotopic to γ such that $\beta(t) = \gamma(t)$ for $t \notin (x,y) + T\mathbb{Z}$, and $\beta_{\mathbb{C}} = \alpha$.

β projects nicely since $\beta_{\mathbb{C}} = \alpha$ is regular, for all t we have $\beta_{\mathbb{C}}(t) = \exp(\alpha(t)) \neq 0$, and by (viii) $D_\beta = \mathrm{Arg}'_\alpha$ has only simple zeroes.

We claim that β is a bending forwards of γ on $[x,y]$. It has regular projection and is smoothly isotopic to γ. Since $\mathrm{Arg}'_\alpha > 0$ on $[a,u]$ we have $\mathrm{Arg}'_\alpha(a) > 0$, $\mathrm{Arg}'_\alpha(u) > 0$. Then since Arg'_α has at most one zero on $[x,a]$, either we have $\mathrm{Arg}'_\alpha > 0$ on $[x,a]$ or there is $w_x \in [x,a)$ such that if $t \in [x,a]$ then $\mathrm{Arg}'_\alpha(t) > 0$ iff $t > w_x$. Similarly either $\mathrm{Arg}'_\alpha > 0$ on $[u,y]$ or there is $w_y \in (u,y]$ such that if $t \in [u,y]$ then $\mathrm{Arg}'_\alpha(t) > 0$ iff $t < w_y$. Thus $\{t \in [x,y] \mid \mathrm{Arg}'_\alpha(t) > 0\}$ is a nonempty interval, i.e. $\{t \in [x,y] \mid D_\beta(t) > 0\}$ is a nonempty interval. As noted we have that if $t \notin (x,y) + T\mathbb{Z}$ then $\beta(t) = \gamma(t)$. Finally if $t \in (x,y)$ then $\mathrm{Arg}'_\alpha(t) > \mathrm{Arg}'_\gamma(t)$, and if $t \in [y-T,x]$ then $\mathrm{Arg}'_\alpha(t) = \mathrm{Arg}'_\gamma(t)$, so for all t we have $D_\beta(t) = \mathrm{Arg}'_\alpha(t) \geqslant \mathrm{Arg}'_\gamma(T)$. Thus β is indeed a bending forwards of γ on $[x,y]$.

Also, we have that if $t \in [a,u]$ then $D_\beta(t) > 0$. Then we claim that if $s \in [a,b]$ is a crossing point of $\beta_{\mathbb{C}}$ such that $D_\beta(s) \leqslant 0$ then s is a crossing point of $\gamma_{\mathbb{C}}$. Suppose first that (s,t) is a crossing pair of $\beta_{\mathbb{C}}$ with $s \in [y,b]$. Then since $\alpha((x,y))$ and $\alpha([y,b])$ are disjoint we must have $t \in [y-T,x] + T\mathbb{Z}$, so that $\beta_{\mathbb{C}}(t) = \alpha(t) = \gamma_{\mathbb{C}}(t)$, and s is a crossing point of $\gamma_{\mathbb{C}}$. But if $s \in [a,y]$ with $D_\beta(s) \leqslant 0$, then we must have $s \geqslant w$ where w is the unique zero of Arg'_α on $[u,y]$, so that s is not a crossing point of $\alpha = \beta_{\mathbb{C}}$. Thus indeed, if $s \in [a,b]$ is a crossing point of $\beta_{\mathbb{C}}$ such that $D_\beta(s) \leqslant 0$ then s is a crossing point of $\gamma_{\mathbb{C}}$. Thus β is a bending forwards of $(\gamma, [a,u])$ on $[x,y]$ without adding crossings to $[a,b]$. □

Now we use this to bend forwards the entire section $[a,b]$.

Proposition 1.20. *Let γ be a smooth knot which projects nicely, let $a < b$ such γ has at most one backwards bend on $[a, b]$, and such that $D_\gamma(t) \leqslant 0$ for $t \in [a, b]$. Then there is a smooth knot β which projects nicely and is a bending forwards of γ such that if $t \in [a, b]$ then $D_\beta(t) > 0$.*

Proof. Since γ has at most one backwards bend on $[a, b]$ we have $b < a + \frac{T}{2}$ and can find an open interval V containing $[a, b]$ with $\sup(V) < \inf(V) + \frac{T}{2}$. Pick $s_0 < s_1 < \ldots < s_n \in [a, b]$ with $s_0 = a$, $s_n = b$ such that for each $i = 0 \ldots (n-1)$, $\gamma_{\mathbb{C}}$ has at most one crossing point on $[s_i, s_{i+1}]$. Set $s_{-1} = \inf(V)$, $s_{n+1} = \sup(V)$.

We claim by induction on i that for each $i = 1 \ldots n$ there is a smooth knot β_i which projects nicely and is a bending forwards of $(\gamma, [s_0, s_i])$ on some compact interval I_i with $I_i \subseteq (s_{-1}, s_{i+1})$ without adding crossings to $[a, b]$. For $i = 1$ this follows from proposition A.1.19.

Suppose true for $i < n$, and let β_i be a smooth knot which projects nicely and is a bending forwards $(\beta_i, [s_0, s_i])$ of γ on some compact interval I_i with $I_i \subseteq (s_{-1}, s_{i+1})$ without adding crossings to $[a, b]$. Let $c = \sup D^{\geqslant 0}_{\beta_i}(I_i) \in (s_i, s_{i+1})$. Then if $t \in [c, b]$ then $D_\beta(t) \leqslant 0$, and so if $t \in [c, b]$ is a crossing point of β_i then since β_i is a bending forwards of γ without adding crossings to $[a, b]$, we must have that t is a crossing point of $\gamma_{\mathbb{C}}$. In particular β_i has at most one crossing point on $[c, s_{i+1}]$. Also, we have $a = s_0 \in D{>}\,0_\beta(I_i)$ so that $D{>}\,0_\beta(I_i) \not\subseteq \mathrm{Int}(D^{\leqslant 0}_\gamma([a, b])) = (a, b)$, so that by proposition A.1.9 β_i has at most one backwards bend on $[a, b]$ and thus at most one backwards bend on $[c, b]$. Thus by proposition A.1.19 there is a smooth knot β_{i+1} which projects nicely and is a bending forwards of $(\beta_i, [c, s_{i+1}])$ on some compact interval $J \subseteq (s_{-1}, s_{i+2})$ without adding crossings to $[c, b]$. Let $I_{i+1} = I_i \cup J \subseteq (s_{-1}, s_{i+2})$.

We will argue that β_{i+1} is a bending forwards of $(\gamma, [s_0, s_{i+1}])$ on I_{i+1} without adding crossings to $[a, b]$ (we already have $I_{i+1} \subseteq (s_{-1}, s_{i+2})$). Since $D_{\beta_{i+1}}(c) > 0$ we must have $c \in J$, and also $c = \sup D^{\geqslant 0}_{\beta_i}(I_i) \in I_i$ so that $c \in J \cap I_i$ with $D_{\beta_i}(c) \geqslant 0$. Also we have $\sup(I_i + 1) - \inf(I_{i+1}) \leqslant s_{n+1} - s_{-1} < \frac{T}{2}$. Thus by proposition A.1.8, β_{i+1} is a bending

forwards of γ on I_{i+1}. We have that if $t \in [s_0, s_{i+1}]$ then $D_{\beta_{i+1}}(t) > 0$. If $t \in [a, b]$ is a crossing point of β_{i+1} with $D_{\beta_{i+1}}(t) \leqslant 0$ then since β_{i+1} is a bending forwards of β_i without adding crossings to $[a, b]$ we have that t is a crossing point of β_i, and then since $D_{\beta_i}(t) \leqslant D_{\beta_{i+1}}(t) \leqslant 0$ and β_i is a bending forwards of γ without adding crossings to $[a, b]$ we have that t is a crossing point of γ. Thus indeed β_{i+1} projects nicely and is a bending forwards of $(\gamma, [s_0, s_{i+1}])$ on I_{i+1} without adding crossings to $[a, b]$.

Thus by induction for each i we can indeed find a smooth knot β_i which projects nicely and is a bending forwards of $(\gamma, [s_0, s_i])$ on some compact interval I_i with $I_i \subseteq (s_{-1}, s_{i+1})$ without adding crossings to $[a, b]$. But taking $i = n$ we have that there is a smooth knot β_n which projects nicely and is a bending forwards of γ with $D_\beta(t) > 0$ for all $t \in [a, b]$, as required. □

Now we can finally prove the Bending Forward proposition.

Proposition 1.21 (Bending forward one troublesome part)**.** *Let γ be a smooth knot which projects nicely. Let U be an open interval in $\mathbb{R}$ such that γ has at most one backwards bend on U. Then there is a smooth knot β which projects nicely and is a bending forwards of γ such that if $t \in U$ then $D_\beta(t) > 0$.*

Proof. Let γ be a smooth knot which projects nicely and let U be an open interval such that γ has at most one backwards bend on U. If $D_\gamma^{\leqslant 0}(U) = \varnothing$ we are done, so we may assume $D_\gamma^{\leqslant 0}(U) \neq \varnothing$. Then if $t \in D_\gamma^{\leqslant 0}(U)$ then $D_\gamma(t) \leqslant 0$ and D_γ has only simple zeroes, so we have some proper interval containing t contained in$D_\gamma^{\leqslant 0}(U)$. Thus $D^{\leqslant 0})\gamma(\overline{U})$ is a proper interval, and we can find $a < b$ with $D_\gamma^{\leqslant 0}(\overline{U}) = [a, b]$. Then by proposition A.1.20 there is a smooth knot β which projects nicely and is a bending forwards of γ such that if $t \in [a, b]$ then $D_\beta(t) > 0$; but thus if $t \in U$ then $D_\beta(t) > 0$, and we are done. □

Before concluding, one final definition is helpful.

Definition 1.22. Let γ be a smooth knot which projects nicely. Say that γ is **($\leqslant n$)-troublesome** where $n \in \mathbb{N}$ if we can write $\pi(D_\gamma^{\leqslant 0}([0,T])) \subseteq \bigcup_{i\in I} \pi(U_i)$ where $|I| = n$ and for each i, γ has at most one backwards bend on U_i.

Thus if γ is ($\leqslant n$)-troublesome, it is also ($\leqslant m$)-troublesome for every $m \geqslant n$. $D_\gamma^{\leqslant 0}([0,T])$ is covered by open intervals U_i on which γ has at most one backwards bend, as discussed after definition A.1.6. Thus $\pi(D_\gamma^{\leqslant 0}([0,T]))$ is covered by sets $\pi(U_i)$, and then by compactness of $\pi(D_\gamma^{\leqslant 0}([0,T]))$ is covered by finitely many such sets (since π is an open map). Thus every smooth knot which projects nicely is n-troublesome for some $n \in \mathbb{N}$.

Now we put proposition A.1.21 together with proposition A.1.9.

Proposition 1.23 (Decreasing trouble)**.** *Let γ be a smooth knot which is ($\leqslant n$)-troublesome. Then γ is smoothly isotopic to a smooth knot β which is ($\leqslant n-1$)-troublesome.*

Proof. Suppose γ is a smooth knot which is ($\leqslant n$)-troublesome. Write

$$\pi(D_\gamma^{\leqslant 0}([0,T])) \subseteq \bigcup_{i\in I} \pi(U_i)$$

with $|I| = n$ and where for each i, γ has at most one backwards bend on U_i. If for some $i \neq j$ we have $\pi(D_\gamma^{\leqslant 0}(U_i)) \subseteq \pi(U_j)$ then we can discard U_i and obtain that γ is ($\leqslant n-1$)-troublesome. Thus we may assume that for all $i \neq j$, $\pi(D_\gamma^{\leqslant 0}(U_i)) \nsubseteq \pi(U_j)$. But then for all $i \neq j$, $D_\gamma^{\leqslant 0}(U_i) \nsubseteq \pi^{-1}(\pi(U_j)) = U_j + T\mathbb{Z}$.

Picking $i \in I$, by proposition A.1.21 we can let β be a smooth knot which projects nicely and is a bending forwards of γ on some compact interval J and such that if $t \in U_i$ then $D_\beta(t) > 0$.

Letting $D_\beta^{>0}(J) = \{t \in J \mid D_\beta(t) > 0\}$ we have that $D_\gamma^{\leqslant 0}(U_i) \subseteq D_\beta^{>0}(J)$, so if $j \neq i$ we have $D_\beta^{>0}(J) \nsubseteq U_j + T\mathbb{Z}$, so $D_\beta^{>0}(J) \nsubseteq \mathrm{Int}(D_\gamma^{\leqslant 0}(\overline{U_j})) + T\mathbb{Z}$. Thus by proposition A.1.9 β has at most one backwards bend on $\overline{U_j}$, so has at most one backwards bend on U_j.

Taking $L = \{i \in I \mid D_\beta^{\leqslant 0}(U_i) \neq \varnothing\}$, we have $|L| < |I|$. Suppose that $t \in D_\beta^{\leqslant 0}([0,T])$, so $D_\beta(t) \leqslant 0$. Then $D_\gamma(t) \leqslant 0$, so $\pi(t) \in \pi(U_i)$ for some i, i.e. there is $t' \in U_i$ with $t \equiv t'$. Thus $i \in L$, and $t' \in U_i$, so $t \in \pi(U_i) \subseteq \bigcup_{j\in L} \pi(U_j)$. Thus $\pi(D_\beta^{\leqslant 0}([0,T])) \subseteq \bigcup_{j\in L} \pi(U_j)$, as required. □

Finally we have our desired conclusion.

Lemma 1.24 (Alexander's lemma)**.** *Let γ be a smooth knot. Then there is a smooth knot β which is smoothly isotopic to γ such that β has regular projection avoiding 0, and we have $D_\beta(t) > 0$ for all t.*

Proof. By proposition A.1.23 and induction, if γ is $(\leqslant n)$-troublesome for any n then γ is smoothly isotopic to a smooth knot β which is $(\leqslant 0)$-troublesome, and thus with $D_\beta(t) > 0$ for all t.

Now let γ be any smooth knot. By proposition A.1.10 together with proposition A.1.5 the set of smooth knots which project nicely is dense in the set of smooth knots. Thus by theorem A.3.8, every smooth knot is isotopic to a smooth knot which projects nicely. Thus we can find a smooth knot γ^* which projects nicely which is smoothly isotopic to γ.

As discussed after definition A.1.22, γ^* is n-troublesome for some n. But then as initially noted, γ^* is smoothly isotopic to a smooth knot β with $D_\beta(t) > 0$ for all t. Since γ is thus smoothly isotopic to β we are done. □

2 Smooth and periodic functions

If U is an open subset of $\mathbb{R}^n$ then a smooth function $f : U \to \mathbb{R}^m$ is one which has continuous partial derivatives of all orders on U. Then identity functions are smooth, linear and bilinear maps are smooth, products and sums of smooth functions are smooth, and compositions of smooth functions are smooth.

Then for arbitrary subsets A of $\mathbb{R}^n$, smooth functions on A are defined to be those which have smooth extensions in the neighbourhood of each point.

Definition 2.1. Let $A \subseteq \mathbb{R}^n$. A function $f : A \to \mathbb{R}^m$ is **smooth** if for every $a \in A$ there is a neighbourhood U of a in $\mathbb{R}^n$ and a smooth function $\tilde{f} : U \to \mathbb{R}^m$ such that $\tilde{f}|_{A\cap U} = f|_{A\cap U}$.

If $A \subseteq \mathbb{R}$ and B is a linear subspace of $\mathbb{R}^n$ we let $C^\infty(A, B)$ denote the vector space of smooth functions f from A to $\mathbb{R}^n$ such that $\text{Image}(f) \subseteq B$. Again, identity functions on arbitrary subsets of $\mathbb{R}$ are smooth, products and sums of smooth functions are smooth, and compositions of smooth functions are smooth. Also, functions that are locally smooth are smooth: if $A \subseteq \mathbb{R}$ and $A = \bigcup_{i\in I} U_i$ with each U_i an open subset of A, and $f : A \to \mathbb{R}^n$ has that $f|_{A\cap U_i}$ is smooth for all $i \in I$, then f is smooth.

If A is a compact subset of $\mathbb{R}$ and V a subspace of $\mathbb{R}^n$ then we equip $C^\infty(A, V)$ with the norm

$$\|\gamma\|_{C^1} = \sup_{t\in A}\|\gamma(t)\| + \sup_{t\in A}\|\gamma'(t)\|$$

(similar to the norm on $C^\infty_T(\mathbb{R}, V)$).

The key difference between smooth functions and real analytic functions is that smooth functions are not rigid: they can be going "along one path" at one point, and then smoothly change to going "along another path". This difference is exemplified by the existence of smooth cutoff functions.

Proposition 2.2. *There exists a smooth function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = 0$ for $x \leqslant 0$, $f(x) = 1$ for $x \geqslant 1$, and f is strictly increasing on $(0, 1)$.*

Proof. This is a standard fact, see for instance Hirsch (1976, p.42), Guillemin and Pollack (1974, p.7) or J. Lee (2012, pp.41–42). □

One can use such cutoff functions to glue together other smooth functions and obtain very varied results, often corresponding to intuitive ideas about how functions can be

"bent". Some examples of this are seen in propositions A.2.12 to A.2.15.

Proposition 2.3. *Suppose that $\gamma : \mathbb{R} \to \mathbb{R}^n$ is differentiable, and that I is a proper interval such that there is $u \in \mathbb{R}^n$ such that for all $s \in I$, $\gamma'(s) \cdot u > 0$. Then γ is injective on I.*

Proof. The map $s \mapsto \gamma(s) \cdot u$ has positive derivative on I so is strictly increasing on I, so is injective. Thus $s \mapsto \gamma(s)$ must be injective on I. □

Proposition 2.4. *Suppose $\gamma : \mathbb{R} \to \mathbb{R}^n$ is an immersion. Then γ is locally injective, i.e. for all t there is an open interval U around t such that $\gamma|_U$ is injective.*

Proof. Given $t \in \mathbb{R}$, we have $\gamma'(t) \cdot \gamma'(t) = \|\gamma'(t)\|^2 > 0$ and γ' is continuous so there is an open interval U around t such that if $s \in U$ then $\gamma'(s) \cdot \gamma'(t) > 0$. Thus we are done by proposition A.2.3. □

Proposition 2.5. *Suppose $\gamma : U \to \mathbb{R}^n$ is an immersion with U an open interval, and that $a < b \in U$. If $\gamma|_{[a,b]}$ is injective then there is $c < a$ such that $\gamma|_{(c,b]}$ is injective.*

Proof. We can find $\delta > 0$ such that $\gamma|[a - \delta, a + \delta]$ is injective by proposition A.2.4, and then since $\gamma(a) \notin \gamma([a + \delta, b])$ there is $\epsilon > 0$ such that $B_\epsilon(\gamma(a))$ is disjoint from $\gamma([a + \delta, b])$, and then we can find $\delta' \in (0, \delta)$ such that $\gamma([a - \delta', a + \delta']) \subseteq B_\epsilon(\gamma(b))$. Then if $s \in [a - \delta', b]$ and $t \in [a - \delta', b]$ with $\gamma(s) = \gamma(t)$ then by the choice of δ' we cannot have $t \in [a + \delta, b]$ so must have $t < a + \delta$, but then by the choice of δ we must have $t = s$. Thus indeed taking $c = a - \delta'$ we have that $\gamma|_{[c,b]}$ is injective. □

Proposition 2.6. *Suppose $\gamma : \mathbb{R} \to X$ is T-periodic with X Hausdorff. Then if $A \subseteq \mathbb{R}$ is a saturated closed set then $\gamma(A)$ is closed, and if $U \subseteq \mathbb{R}$ is a saturated open set then $\gamma(U)$ is an open subset of $\gamma(\mathbb{R})$.*

Proof. We do the saturated closed sets first. Let $A \subseteq \mathbb{R}$ be saturated and closed. Then $f(A) = f(A \cap [0, T])$ is a compact set therefore closed (since X is Hausdorff). Now let

U be saturated and open. Then $f(U) = f(\mathbb{R}) \setminus f(U^{\mathsf{c}})$ is an open subset of $f(\mathbb{R})$, since $f(U^{\mathsf{c}})$ is a closed subset of X, and thus of $f(\mathbb{R})$. □

Proposition 2.7. *Suppose γ is a smooth knot. Then if $U \subseteq \mathbb{R}$ is open then $\gamma(U)$ is an open subset of $\gamma(\mathbb{R})$.*

Proof. We have that $\gamma(U) = \gamma(U + T\mathbb{Z})$ which is an open subset of $\gamma(\mathbb{R})$ by proposition A.2.6. □

Proposition 2.8. *Suppose I is a proper interval and $\gamma : I \to \mathbb{R}^n$ is smooth. Suppose that $\sup(I) - \inf(I) > T$ and γ is T-periodic on I, i.e. that if $s, t \in I$ with $s - t \in T\mathbb{Z}$ then $\gamma(s) = \gamma(t)$. Then there is a unique T-periodic map $\overline{\gamma} : \mathbb{R} \to \mathbb{R}^n$ such that $\overline{\gamma}|_I = \gamma$. We call $\overline{\gamma}$ the T-periodic extension of γ to $\mathbb{R}$.*

Proof. Since $\sup(I) - \inf(I) >$, for every $t \in \mathbb{R}$ there is $t' \in I$ with $t \equiv t'$. We define $\overline{\gamma}(t)$ to be $\gamma(t')$ for any such t'. This is well defined since γ is T-periodic, and is smooth since it is smooth on each $\mathrm{Int}(I) + Tk$ for $k \in \mathbb{Z}$, and these are open sets which cover $\mathbb{R}$. □

Proposition 2.9. *Let $\gamma : U \to \mathbb{R}^n$ be a C^1 curve, with U an open interval and $n \geqslant 2$. Then* Image(γ) *has measure zero, and there is no open set $U \subseteq \mathbb{R}^n$ such that $U \subseteq$* Image(γ).

Proof. Cover U by countably many sections $[a_i, b_i]$ on which γ is Lipschitz, i.e. there is $K > 0$ such that $\|\gamma(s) - \gamma(t)\| \leqslant K|s - t|$. There is $L > 0$ such that the measure of a ball of radius ϵ in $\mathbb{R}^n$ is $L\epsilon^n$. Fix i. Given $m \in \mathbb{N}$, we can divide $[a_i, b_i]$ into m sections $[c_0, c_1], \ldots [c_{m-1}, c_m]$, with $c_{j+1} - c_j = \frac{b_i - a_i}{m}$ for each j. Then each $\gamma([c_j, c_{j+1}])$ is compact and therefore measurable, and we have

$$\gamma([c_j, c_{j+1}]) \subseteq B_{\frac{K(b_i - a_i)}{m}}(\gamma(c_j)),$$

so the measure of $\gamma([c_j, c_{j+1}])$ is at most $L(\frac{K(b_i - a_i)}{m})^n$. Thus the measure of $\gamma([a_i, b_i])$ is at most $nL(\frac{K(b_i - a_i)}{m})^n$. As $n \to \infty$ this quantity tends to zero, so the measure of

$\gamma([a_i, b_i])$ is zero. $\gamma(U)$ is a countable union of such sets, so has measure zero. Thus it can contain no open set. □

Proposition 2.10. *The map* $\exp_* : C^\infty_T(\mathbb{R}, \mathbb{C}) \to C^\infty_T(\mathbb{R}, \mathbb{C})$, $\alpha \mapsto \exp \circ \alpha$ *is continuous.*

Proof. Let $\alpha \in C^\infty_T(\mathbb{R}, \mathbb{C})$ and let $\epsilon > 0$. We seek $\delta > 0$ such that $\|\beta - \alpha\|_{C^1} < \delta$ implies $\|\exp \circ \beta - \exp \circ \alpha\|_{C^1} < \epsilon$. WLOG we may assume $\epsilon \leqslant 1$.

Let $K = \max(\|\alpha\|_{C^1}, \|\exp \circ \alpha\|_{C^1}, 1)$. Let $A = \{z \mid |z| \leqslant K + 1\}$. The set A is compact so exp is uniformly continuous on it. Thus we can find $\delta > 0$ such that if $|w - z| < \delta$ then $|\exp(w) - \exp(z)| < \frac{\epsilon}{3K}$. WLOG we may assume $\delta(K + 1) < \frac{\epsilon}{3}$. Then if $\|\beta - \alpha\|_{C^1} < \delta$ we have for any t that $|\exp(\beta(t)) - \exp(\alpha(t))| \leqslant \frac{\epsilon}{3K} \leqslant \frac{\epsilon}{3}$. Thus for any t, $|\exp(\beta(t))| \leqslant |\exp(\alpha(t))| + |\exp(\beta(t)) - \exp(\alpha(t))| \leqslant K + \frac{\epsilon}{3K} < K + 1$. Then we have

$$\begin{aligned}
|(\exp \circ \beta)'(t) - (\exp \circ \alpha)'(t)| &= |\beta'(t) \exp(\beta(t)) - \alpha'(t) \exp(\alpha(t))| \\
&\leqslant |(\beta'(t) - \alpha'(t)) \exp(\beta(t))| + |\alpha'(t)(\exp(\beta(t)) - \exp(\alpha(t)))| \\
&\leqslant \delta(K + 1) + K|\exp(\beta(t)) - \exp(\alpha(t))| \\
&\leqslant \delta(K + 1) + K \cdot \frac{\epsilon}{3K} < \frac{\epsilon}{3} + \frac{\epsilon}{3} = \frac{2\epsilon}{3}.
\end{aligned}$$

Thus $\|\exp \circ \beta - \exp \circ \alpha\|_{C^1} < \frac{\epsilon}{3+} \frac{2\epsilon}{3} = \epsilon$, as required. □

Proposition 2.11. *Let* $\gamma : B \to \mathbb{R}^n$ *be smooth with* $B \subseteq \mathbb{R}$ *a proper interval. Then the map*

$$F_\gamma : B^2 \to \mathbb{R}^n,\ (x, y) \mapsto \begin{cases} \frac{\gamma(x) - \gamma(y)}{x - y} & \text{if } x \neq y \\ \gamma'(x) & \text{if } x = y \end{cases}$$

is smooth.

Proof. First, if e is an endpoint of B then γ has a smooth extension to a neighbourhood of e. Thus γ has a smooth extension $\tilde{\gamma}$ to an open interval U with $B \subseteq U$. Thus WLOG we may assume B is open.

By the integral form of the remainder for Taylor's theorem we have for all $(x, y) \in B^2$ that

$$f(y) = f(x) + (y - x)f'(x) + (y - x)^2 \int_0^1 (1 - s)f''(x + s(y - x))\, ds.$$

Thus we have $F_\gamma = f'(x) + (y - x)\int_0^1 (1 - s)f''(x + s(y - x))\, ds$, which is a smooth function by the ability to differentiate under the integral sign (here we use that B^2 is open, often taken as a prerequisite for differentiating under the integral sign). □

Proposition 2.12. *Let I be a proper interval, and let $a > b$ with $[a, b] \subseteq I$. Let $f, g : I \to \mathbb{R}$ be smooth functions. Then there is a smooth function h such that*

- $h(t) = f(t)$ *for* $t \leqslant a$
- $h(t) = g(t)$ *for* $t \geqslant b$
- $\min(f(t), g(t)) \leqslant h(t) \leqslant \max(f(t), g(t))$ *for* $t \in [a, b]$
- $\min(f(t), g(t)) < h(t) < \max(f(t), g(t))$ *for* $t \in (a, b)$ *such that* $f(t) \neq g(t)$

Proof. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a smooth function with $\phi(t) = 0$ for $t \leqslant a$, $\phi(t) = 1$ for $t \geqslant b$, and $\phi(t) \in (0, 1)$ for $t \in (a, b)$. Then the function $h : t \mapsto f(t) + \phi(t)(g(t) - f(t))$ has the required properties. □

Proposition 2.13. *Let I be a proper interval, and let $a > b$ with $[a, b] \subseteq I$. Let $f, g : I \to \mathbb{R}$ be smooth functions with $f(t) \leqslant g(t)$ for $t \in [a, b]$ and $f'(t), g'(t) > 0$ on $[a, b]$. Then there is a smooth function $h : I \to \mathbb{R}$ such that*

- $h(t) = f(t)$ *for* $t \leqslant a$
- $h(t) = g(t)$ *for* $t \geqslant b$
- $f(t) \leqslant h(t) \leqslant g(t)$ *for* $t \in [a, b]$
- $h'(t) \geqslant \min(f'(t), g'(t))$ *for* $t \in [a, b]$

- $h'(t) > \min(f'(t), g'(t))$ *if* $t \in (a, b)$ *and* $f'(t) \neq g'(t)$

Proof. Take ϕ as in the previous proposition, but with the added requirement that $\phi' \geqslant 0$ on $[a, b]$. Let $h(t) = f(t) + \phi(t)(g(t) - f(t))$ as above. Then if $t \in [a, b]$ we have

$$\begin{aligned} h'(t) &= \phi'(t)(g(t) - f(t)) + \phi(t)(g'(t) - f'(t)) + f'(t) \\ &\geqslant \phi(t)(g'(t) - f'(t)) + f'(t) \end{aligned}$$

since $g(t) \geqslant f(t)$. This proves the proposition since $\phi(t)(g'(t)-f'(t))+f'(t) \geqslant \min(f'(t), g'(t))$, and if $f'(t) \neq g'(t)$ and $t \in (a, b)$ then $\phi(t) \in (0, 1)$ so $\phi(t)(g'(t) - f'(t)) + f'(t) > \min(f'(t), g'(t))$. □

Proposition 2.14. *Let I be a proper interval and $f, g : I \to \mathbb{R}$ smooth. Suppose we have $a < b \in I$ such that $f(t) < 0$ and $g(t) > 0$ for $t \in [a, b]$. Then there is a smooth function $h : I \to \mathbb{R}$ such that*

- $h(t) = f(t)$ *for* $t \leqslant a$
- $h(t) = g(t)$ *for* $t \geqslant b$
- $f(t) < h(t) < g(t)$ *for* $t \in (a, b)$
- h *has a single simple zero in* $[a, b]$

Proof. We can find $c, d > 0$ such that if $t \in [a, b]$ then $f(t) \leqslant -c$ and $g(t) \geqslant d$. Then by proposition A.2.12 there is a smooth function $f_2 : I \to \mathbb{R}$ such that $f_2(t) = f(t)$ for $t \leqslant a$, $f_2(t) = -c$ for $t \geqslant \frac{2a+b}{3}$, and $-c > f_2(t) > f(t)$ for $t \in (a, \frac{2a+b}{3})$. Similarly there is a smooth function $g_2 : I \to \mathbb{R}$ such that $g_2(t) = d$ for $t \leqslant \frac{a+2b}{3}$, $g_2(t) = g(t)$ for $t \geqslant b$, and $d < g_2(t) < g(t)$ for $t \in (\frac{a+2b}{3}, b)$. It follows from these that $f_2(t) > f(t)$ for $t \in (a, b)$ and $g_2(t) < g(t)$ for $t \in (a, b)$.

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a smooth function with $0 \leqslant \phi(t) \leqslant 1$ for all t, $\phi(t) = 0$ for $t \leqslant \frac{2a+b}{3}$, $\phi(t) = 1$ for $t \geqslant \frac{a+2b}{3}$, and $\phi'(t) > 0$ for $t \in (\frac{2a+b}{3}, \frac{a+2b}{3})$. Then we can take

$$h(t) = \phi(t)(g_2(t) - f_2(t)) + f_2(t).$$

This is smooth and satisfies $h(t) = f_2(t)$ for $t \leqslant \frac{2a+b}{3}$ and $h(t) = g_2(t)$ for $t \geqslant \frac{a+2b}{3}$. Thus we have $h(t) = f(t)$ for $t \leqslant a$ and $h(t) = g(t)$ for $t \geqslant b$. If $t \in (a, b)$ we have $f(t) < f_2(t) \leqslant h(t) \leqslant g_2(t) < g(t)$, so $f(t) < h(t) < g(t)$.

Since $h(t) = f_2(t) \leqslant -c$ for $t \in [a, \frac{2a+b}{3}]$, h has no zeroes on $[a, \frac{2a+b}{3}]$. Similarly we have $h(t) = g_2(t) \geqslant d$ for $t \in [\frac{a+2b}{3}, b]$ so h has no zeroes on $[\frac{a+2b}{3}, b]$. Finally on $(\frac{2a+b}{3}, \frac{a+2b}{3})$ we have $f_2(t) = -c$ and $g_2(t) = c$ so

$$h'(t) = \phi'(t)(d - (-c)) > 0.$$

Thus h has a unique simple zero on $[\frac{2a+b}{3}, \frac{a+2b}{3}]$, so has a unique simple zero on $[a, b]$. □

Proposition 2.15. *Let I be a proper interval and $f, g : I \to \mathbb{R}$ smooth. Let $a < b \in I$ such that $g(a) \geqslant f(a)$ and if $t \in [a, b]$ then $f'(t) < 0$, $g'(t) > 0$. Then there is a smooth function $h : I \to \mathbb{R}$ such that*

- *$h(t) = f(t)$ for $t \leqslant a$*
- *$h(t) = g(t)$ for $t \geqslant b$*
- *$f(t) < h(t) \leqslant g(t)$ for $t \in (a, b)$*
- *$f'(t) < h'(t)$ for $t \in (a, b]$*
- *h' has a single simple zero on $[a, b]$*

Proof. By proposition A.2.14 we can find a smooth function $\theta : I \to \mathbb{R}$ such that $\theta(t) = f'(t)$ for $t \leqslant a$, $\theta(t) = g'(t)$ for $t \geqslant \frac{a+b}{2}$, $f'(t) < \theta(t) < g'(t)$ for $t \in (a, \frac{a+b}{2})$, and with θ having a single simple zero in $[a, \frac{a+b}{2}]$.

Let $\phi : \mathbb{R} \to \mathbb{R}$ be any smooth non zero non negative function such that if $\phi(t) > 0$ then $t \in (\frac{a+b}{2}, b)$. For $k \geqslant 0$ let

$$\psi_k : I \to \mathbb{R},\ t \mapsto \begin{cases} \int_a^t (\theta(t) + k\phi(t))\ dt + f(a) & \text{if } t \geqslant a \\ f(a) - \int_t^a (\theta(t) + k\phi(t))\ dt & \text{if } t \leqslant a. \end{cases}$$

By the fundamental theorem of calculus, for each k this is smooth with $\psi_k'(t) = \theta(t) + k\phi(t)$. Thus we have $\psi_k'(t) > f'(t)$ for $t \in (a, b]$, so $\psi_k(t) > f(t)$ for $t \in (a, b]$. If $t \in [a, \frac{a+b}{2}]$ then $\psi_k'(t) = \theta(t) \leqslant g'(t)$, so if $t \in [a, \frac{a+b}{2}]$ then $\psi_k(t) \leqslant f(a) + \int_0^1 g'(t) \leqslant g(t)$. Also if $t \in (\frac{a+b}{2}, b)$ then $\psi_k'(t) = \theta(t) + k\phi(t) \geqslant \theta(t) = g'(t)$. Additionally if $t \in [\frac{a+b}{2}, b]$ then $\psi_k'(t) \geqslant \theta(t) \geqslant g'(t) > 0$, and if $t \in [a, \frac{a+b}{2}]$ then $\psi_k(t) = \theta(t)$, so ψ_k' has a single simple zero on $[a, b]$.

We have $\psi_k(a) = f(a)$, and $\phi_0(b) = f(a) + \int_a^b \theta(t)\ dt \leqslant g(a) + \int_a^b g(t)\ dt = g(b)$. Then

$$\psi_k(b) = \int_a^b \theta(t) + k\phi(t)\ dt + f(a) = k\int_a^b \phi(t)\ dt + \int_a^b \theta(t)\ dt + f(a)$$

which is a continuous function of k which tends to ∞ as k tends to ∞ (since $\int_a^b \phi(t)\, dt > 0$), so there is $k \geqslant 0$ with $\psi_k(b) = g(b)$. Thus $\psi_k(t) = g(t)$ for $t \geqslant b$. Then for this k for all $t \in [\frac{a+b}{2}, b]$ we have $\psi_k'(t) \geqslant g'(t)$, so for any such t we must have $\psi_k(t) \leqslant g(t)$. Also it was noted above that if $t \in [a, \frac{a+b}{2}]$ then $\psi_k(t) \leqslant g(t)$, so for all $t \in [a, b]$ we have $\psi_k(t) \leqslant g(t)$.

Then taking this value of k we have that $\psi_k : I \to \mathbb{R}$ is smooth, that $\psi_k(t) = f(t)$ for $t \leqslant a$, that $\psi_k(t) = g(t)$ for $t \geqslant b$, that $f(t) < \psi_k(t) \leqslant g(t)$ for $t \in (a, b)$, that $f'(t) < \psi_k'(t)$ for $t \in (a, b)$, and that ψ_k' has a single simple zero in $[a, b]$. □

3 Smooth knots and projections

To cover at the same time certain facts both about smooth knots and about projections of smooth knots onto a plane, here V will be used to denote either R^3 or the subspace $\mathbb{R}^2 \times \{0\}$ of $\mathbb{R}^3$, which can be identified with $\mathbb{R}^2$ (and with $\mathbb{C}$). One can derive facts about projection onto a general plane $P \subseteq \mathbb{R}^3$ from facts about projection onto $\mathbb{R}^2 \subseteq \mathbb{R}^3$ by a rotation (and translation) of space.

If $\gamma : A \to \mathbb{R}^n$ is a curve with A a compact interval, we can define the notions of crossing pair and crossing point for γ in a similar way as for T-periodic curves. We say that $(s,t) \in A^2$ is a **crossing pair** of γ if $s \neq t$ and $\gamma(s) = \gamma(t)$. We say that $s \in A$ is a **crossing point** of γ if (s,t) is a crossing pair of γ for some $t \in A$.

Proposition 3.1. *The relation $\sim$ of there being a smooth isotopy from β to γ is an equivalence relation on knots.*

Proof. If γ is a smooth knot then the map $H : [0,1] \times \mathbb{R}, (s,t) \mapsto \gamma(t)$ is smooth – it is $\gamma \circ \pi_2$ where $\pi_2 : [0,1] \times \mathbb{R} \to \mathbb{R}$ is the second projection, a restriction of a linear map. For all s we have $H_s = \gamma$ which is a smooth knot, and $H_0 = H_1 = \gamma$, so $\sim$ is reflexive.

If β and γ are smooth knots, and H is a smooth isotopy from β to γ, then $\tilde{H} : (s,t) \mapsto H(1-s,t)$ is a smooth isotopy from γ to β, so $\sim$ is symmmetric.

Now suppose H is a smooth isotopy from α to β, and J is a smooth isotopy from β to γ. Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth bump function as described in proposition A.2.2. Let $0 < \epsilon < \frac{1}{2}$ and let $g(t) = f(\frac{1}{1-2\epsilon}(t-\epsilon))$. Then g is smooth, and we have $g(t) = 0$ for $t \leqslant \epsilon$, $g(t) = 1$ for $t \geqslant 1-\epsilon$, and g is strictly increasing on $(\epsilon, 1-\epsilon)$.

Let

$$G(s,t) = \begin{cases} H(g(2s),t) & \text{if } s \leqslant \frac{1}{2} \\ J(g(2s-1),t) & \text{if } s \geqslant \frac{1}{2}. \end{cases}$$

This is well defined since $H(1,t) = \beta(t) = J(0,t)$. We have that $G|_{[0,\frac{1}{2})\times\mathbb{R}} : (s,t) \mapsto H(g(2s),t)$ is smooth, and $G|_{(\frac{1}{2},1]\times\mathbb{R}} : (s,t) \mapsto J(g(2s-1),t)$ is smooth. But also if

$s \in (\frac{1-\epsilon}{2}, \frac{1}{2}]$ then we have $g(2s) = 1$, so $H(g(2s), t) = \beta(t)$, and similarly if $s \in [\frac{1}{2}, \frac{1+\epsilon}{2})$ then $g(2s-1) = 0$ so $J(g(2s-1)) = \beta(t)$. Thus we have that $G|_{(\frac{1-\epsilon}{2}, \frac{1+\epsilon}{2})\times\mathbb{R}} : (s,t) \mapsto \beta(t)$ is smooth. Thus G is locally smooth (smooth on each of a collection of open subsets of $[0,1] \times \mathbb{R}$, which collectively cover $[0,1] \times \mathbb{R}$) so is smooth. Finally we have $G_0 = \alpha$ and $G_1 = \gamma$, so $\alpha \sim \gamma$. Thus $\sim$ is transitive. □

Proposition 3.2. *The set of elements of $C^\infty_T(\mathbb{R}, V)$ that are immersions is open.*

Proof. Let $\gamma \in C^\infty_T(\mathbb{R}, V)$ be a smooth immersion. Since $\gamma' \neq 0$ and γ is periodic we can find $\delta > 0$ such that $\|\gamma'(t)\| > \delta$ for all t. Then if $\beta \in C^\infty_T(\mathbb{R}, V)$ and $\|\beta - \gamma\|_{C^1} < \delta$ we have $\|\beta'(t)\| > 0$ for all t, so β is an immersion. □

We denote the set of elements of $C^\infty_T(\mathbb{R}, V)$ that are immersions by $C^{\mathrm{Imm}}_T(\mathbb{R}, V)$.

Proposition 3.3. *Let*

$$\begin{aligned}&\mathrm{NotCrossPair}_V\\ &= \{(\gamma, s, t) \in C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}^2 \mid (s,t) \textit{ is not a crossing pair of } \gamma\}\\ &= \{(\gamma, s, t) \in C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}^2 \mid s \equiv t \textit{ or } \gamma(s) = \gamma(t)\}.\end{aligned}$$

Then $\mathrm{NotCrossPair}_V$ *is an open subset of* $C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}^2$.

Proof. Let (γ, s, t) be an element of $\mathrm{NotCrossPair}_V$.

Suppose we have $s \equiv t$, with $t = s + kT$, $k \in \mathbb{Z}$. Then $\gamma'(s) \neq 0$, and we can find $\delta > 0$ and an open interval U around s such that $\gamma'(x) \cdot \gamma'(s) > \frac{\|\gamma'(s)\|^2}{2}$ on U. Then if $\|\beta - \gamma\|_{C^1} < \frac{\|\gamma'(s)\|^2}{2}$ then $\gamma'(x) \cdot \gamma'(s) > 0$ so by proposition A.2.3 we have that β is injective on U. Then if we have $x \in U$ and $y \in U + kT$ with $\beta(x) = \beta(y)$, then we have $\beta(x) = \beta(y - kT)$ with $y - kT \in U$ so $x = y - kT$ so $x \equiv y$. Thus β has no crossing pairs on $U \times U + kT$. Thus $B_\delta(\gamma) \times U \times (U + kT)$ is an open neighbourhood of (γ, s, t) contained in $\mathrm{NotCrossPair}_V$.

Now suppose we have $\gamma(s) \neq \gamma(t)$. We can find $\delta > 0$ such that $B_{2\delta}(\gamma(s))$ and $B_{2\delta}(\gamma(t))$ are disjoint, and we can find open intervals U around s and V around t such that $\gamma(U) \subseteq B_\delta(\gamma(s))$ and $\gamma(V) \subseteq B_\delta(\gamma(t))$. Then if $\|\beta - \gamma\|_{C^1} < \delta$ then $\beta(U) \subseteq B_{2\delta}(\gamma(s))$ and $\beta(V) \subseteq B_{2\delta}(\gamma(t))$, so $\beta(U)$ and $\beta(V)$ are disjoint, so β has no crossing pairs on $U \times V$. Thus $B_\delta(\gamma) \times U \times V$ is an open neighbourhood of (γ, s, t) contained in $\mathrm{NotCrossPair}_V$. □

Proposition 3.4. *Let*

$$\mathrm{NotCrossPoint}_V = \{(\gamma, s) \in C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R} \mid s \textit{ is not a crossing point of } \gamma\}.$$

Then $\mathrm{NotCrossPoint}_V$ *is an open subset of* $C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}$.

Proof. Note that we have

$$\begin{aligned}\mathrm{NotCrossPoint}_V &= \{(\gamma, s) \mid \text{for all } t \in \mathbb{R},\ (\gamma, s, t) \in \mathrm{NotCrossPair}_V\} \\ &= \{(\gamma, s) \mid \text{for all } t \in [0, T],\ (\gamma, s, t) \in \mathrm{NotCrossPair}_V\}.\end{aligned}$$

Thus if $(\gamma, s) \in \mathrm{NotCrossPoint}_V$ then we have $\{\gamma\} \times \{s\} \times [0, T] \subseteq \mathrm{NotCrossPair}_V$, so since $[0, T]$ is compact and $\mathrm{NotCrossPair}_V$ is open there is an open neighbourhood $U \subseteq C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}$ of $\{(\gamma, s)\}$ such that $U \times [0, T] \subseteq \mathrm{NotCrossPair}_V$. Thus $U \subseteq \mathrm{NotCrossPoint}_V$. □

Proposition 3.5. *Let* $\gamma \in C^{\mathrm{Imm}}_T(\mathbb{R}, V)$. *Then the set of crossing pairs of* γ *is a closed set, as is the set of crossing points.*

Proof. Let $\iota^2_\gamma : \mathbb{R}^2 \to C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}^2,\ (s, t) \mapsto (\gamma, s, t)$, and $\iota^1_\gamma : \mathbb{R} \to C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R},\ s \mapsto (\gamma, s)$, which are continuous. The set of crossing pairs of γ is $((\iota^2_\gamma)^{-1}(\mathrm{NotCrossPair}_V))^{\mathsf{c}}$, which is closed since $\mathrm{NotCrossPair}_V$ is open. Similarly the set of crossing points of γ is $((\iota^1_\gamma)^{-1}(\mathrm{NotCrossPoint}_V))^{\mathsf{c}}$, which is closed. □

Theorem 3.6. *The set of smooth knots is an open subset of* $C^\infty_T(\mathbb{R}, \mathbb{R}^3)$.

Proof. Note that the set of smooth knots is the set

$$\begin{aligned}&\{\gamma \in C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^3) \mid \text{for all s} \in \mathbb{R},\ (\gamma, s) \in \mathrm{NotCrossPoint}_V\}\\ =\ &\{\gamma \in C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^3) \mid \text{for all s} \in [0, T],\ (\gamma, s) \in \mathrm{NotCrossPoint}_V\},\end{aligned}$$

and $\mathrm{NotCrossPoint}_V$ is open by proposition A.3.4. Thus if γ is a smooth knot then we have $\{\gamma\} \times [0, T] \subseteq \mathrm{NotCrossPoint}_V$, so there is an open neighbourhood $W \subseteq C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^3)$ of γ such that $W \times [0, T] \subseteq \mathrm{NotCrossPoint}_V$, so W consists only of smooth knots, and is open in $C^\infty_T(\mathbb{R}, \mathbb{R}^3)$ since $C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^3)$ is. □

By a very similar argument we obtain:

Proposition 3.7. *Let A be a compact proper interval. Then the set of $\alpha \in C^\infty(A, V)$ which are injective smooth immersions is an open set.*

Theorem 3.8. *Smooth isotopy classes of knots are open subsets of* $C^\infty_T(\mathbb{R}, \mathbb{R}^3)$.

Proof. It suffices to show that for all γ, there is $\epsilon > 0$ such that if $\|\beta - \gamma\|_{C^1} < \epsilon$ then β is smoothly isotopic to γ. By theorem A.3.6 there is $\epsilon > 0$ such that if $\|\beta - \gamma\|_{C^1} < \epsilon$ then β is a smooth knot. Then if $\|\beta - \gamma\|_{C^1} < \epsilon$, we can construct a smooth isotopy from γ to β. Indeed, let

$$H(s, t) = \gamma(t) + s(\beta(t) - \gamma(t)).$$

This is visibly smooth. For all s and t we have

$$\|H_s - \gamma\|_{C^1} = \|s(\beta(t) - \gamma(t))\|_{C^1} \leqslant \|\beta - \gamma\|_{C^1} < \epsilon$$

so H_s is a smooth knot for all s. Since $H_0 = \gamma$ and $H_1 = \beta$, H is the required smooth isotopy. □

Proposition 3.9. *Let $u, v \in \mathbb{R}^n$ be non zero and non parallel, $n \in \mathbb{N}^\times$. Then there is $\delta > 0$ such that for any smooth curves $\gamma : I \to \mathbb{R}^n$, $\beta : J \to \mathbb{R}^n$ with $\|\gamma'(x) - u\| < \delta$ for $x \in I$ and $\|\beta'(y) - v\| < \delta$ for $y \in J$, where I and J are proper intervals, then there is at most one pair $(x, y) \in I \times J$ with $\gamma(x) = \beta(y)$, and if there is such a pair then $\gamma'(s)$ and $\beta'(t)$ are non zero and non parallel.*

Proof. We prove a little intermediary result, which is that if $u, v \in \mathbb{R}^n$ are non zero and non parallel then there is $\delta > 0$ such that if $x \in B_\delta(u)$, $y \in B_\delta(v)$ then x, y are non zero and non parallel. Indeed since $\frac{u}{\|u\|} \neq \frac{v}{\|v\|}$ there are open subsets U and V of S^{n-1} which are neighbourhoods of $\frac{u}{\|u\|}$ and $\frac{v}{\|v\|}$ respectively, such that $U \cap V = \varnothing$. Then $U' = \{x \neq 0 \mid \frac{x}{\|x\|} \in U\}$ and $V' = \{y \neq 0 \mid \frac{y}{\|y\|} \in V\}$ are disjoint open neighbourhoods of u and v respectively. Thus we can pick $\delta > 0$ such that $B_\delta(u) \subseteq U'$ and $B_\delta(v) \subseteq V'$, and the claim is proved.

WLOG we may assume that this δ is small enough that if $\|w - u\| < \delta$ then $w \cdot u > 0$, and that if $\|z - v\| < \delta$ then $z \cdot v > 0$. Suppose that we have curves $\gamma : I \to \mathbb{R}^n$, $\beta : J \to \mathbb{R}^n$ with I and J proper intervals, and $s \in I$ and $t \in J$ such that $\gamma(s) = \beta(t)$, and such that $\|\gamma'(x) - u\| < \delta$ for $x \in I$ and $\|\beta'(y) - v\| < \delta$ for $y \in J$. Then by the choice of δ we have $\gamma'(x) \cdot u > 0$ for $x \in I$ and $\beta'(y) \cdot v > 0$ for $t \in J$ so by proposition A.2.3 we have that γ is injective on I and β is injective on J.

If there are no pairs $(s, t) \in I \times J$ with $\gamma(x) = \beta(y)$ then we are done, so suppose we have such a pair (s, t). We have $\|\gamma'(s) - u\| < \delta$ and $\|\beta'(t) - v\| < \delta$ so by the choice of δ we have that $\gamma'(s)$ and $\beta'(t)$ are non zero and non parallel. Thus we need only show that (s, t) is the only pair in $(x, y) \in I \times J$ such that $\gamma(x) = \beta(y)$. Since γ is injective on I, we have if $x \in I \setminus \{s\}$ that $\gamma(x) \neq \gamma(s) = \beta(t)$. Similarly if $y \in J \setminus \{t\}$ then $\beta(y) \neq \gamma(s)$. Thus we need only show that if $x \in I \setminus \{s\}$ and $y \in J \setminus \{t\}$ then $\gamma(x) \neq \beta(y)$.

If $f : x \mapsto \gamma(x) - ux$ then $\|f'(x)\| = \|\gamma'(x) - u\| < \delta$ for $x \in I$ so $\|f(x) - f(s)\| < \delta|x - s|$ for $x \neq s$, or in other words $\|\gamma(x) - \gamma(s) - u(x - s)\| < \delta|x - s|$, so $\|\frac{\gamma(x) - \gamma(s)}{x - s} - u\| < \delta$. Similarly if $y \in J$ and $y \neq t$ then $\|\frac{\beta(y) - \beta(t)}{y - t} - v\| < \delta$. Thus by the choice of δ, we have

that $\frac{\gamma(x)-\gamma(s)}{x-s}$ and $\frac{\beta(t)-\beta(y)}{t-y}$ are non zero and non parallel. Thus $\gamma(x)-\gamma(s) \neq \beta(t)-\beta(y)$, so since $\gamma(s) = \beta(t)$ we must have $\gamma(x) \neq \beta(y)$, as required. □

Theorem 3.10. *Let $\gamma \in C_T^{\mathrm{Imm}}(\mathbb{R}, V)$, and suppose that γ has only transverse crossings. Then the set of crossing pairs of γ is a closed isolated set, and has finite intersection with any compact set. The set of crossing points of γ also has finite intersection with any compact set.*

Proof. Let C be the set of crossing pairs of γ. This is closed by proposition A.3.5. We show it is isolated. Suppose $s \not\equiv t$ and $\gamma(s) = \gamma(t)$, so $\gamma'(s)$ and $\gamma'(t)$ are not parallel. By proposition A.3.9 there is $\delta > 0$ such that if we have curves $\alpha : I \to \mathbb{R}^3$, $\beta : J \to \mathbb{R}^3$ with I and J proper intervals with $\|\alpha'(x) - \beta'(s)\| < \delta$ for $x \in I$ and $\|\beta'(y) - \beta'(t)\| < \delta$ for $y \in J$, then there is at most one pair $(x, y) \in I \times J$ with $\alpha(x) = \beta(y)$, and if there is such a pair then $\gamma'(s)$ and $\beta'(t)$ are non zero and non parallel. But by continuity of γ' at s and at t we can find open intervals U around s and W around t such that if $x \in U$ then $\|\gamma'(x) - \gamma'(s)\| < \delta$ and if $y \in W$ then $\|\gamma'(y) - \gamma'(t)\| < \delta$. Thus there is at most one pair $(x, y) \in U \times W$ with $\gamma(x) = \gamma(y)$, i.e. $C \cap (U \times W) = \{(s, t)\}$, so C is isolated.

Then since C is closed, if A is compact then $A \cap C$ is compact, and thus both compact and isolated, and so finite. If D is the set of crossing points of γ then if $B \subseteq \mathbb{R}$ is compact then we have that $D \cap B = \pi_1(C \cap (B \times [0, T]))$, where π_1 is projection onto the first co-ordinate. Thus $D \cap B$ is indeed finite. □

By a very similar argument we obtain:

Proposition 3.11. *Let A be a compact proper interval, $\gamma : A \to V$ a smooth immersion such that if $s \neq t$ with $\gamma(s) = \gamma(t)$ then $\gamma'(s)$ and $\gamma'(t)$ are not parallel. Then the set of crossing pairs is a closed and isolated set, and thus a compact and isolated set, and thus a finite set. The set of crossing points of γ is also finite.*

Definition 3.12. If $\gamma \in C_T^{\infty}(\mathbb{R}, V)$, say that (r, s, t) is a **triple crossing** of γ if we have

$r \not\equiv s$, $r \not\equiv t$, $s \not\equiv t$, and $\gamma(r) = \gamma(s) = \gamma(t)$. Say that $(s,t) \in \mathbb{R}^2$ is **at most a double crossing** of γ if for all $r \in \mathbb{R}$ we have that (r,s,t) is not a triple crossing of γ.

It is immediate that (r,s,t) is a triple crossing of γ iff (r,s), (r,t) and (s,t) are crossing pairs of γ, and that if (s,t) is a crossing pair of γ then it is a double crossing iff it is at most a double crossing by the above definition.

Proposition 3.13. *Let*

$$\text{NotCrossTriple}_V =$$
$$\{(\gamma, r, s, t) \in C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^3 \mid (r,s,t) \textit{ is not a triple crossing of } \gamma\}.$$

Then NotCrossTriple$_V$ is an open subset of $C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^3$.

Proof. Let

$$\pi_{1,2} : C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^3 \to C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^2,\ (\gamma, r, s, t) \mapsto (\gamma, r, s)$$
$$\pi_{1,3} : C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^3 \to C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^2,\ (\gamma, r, s, t) \mapsto (\gamma, r, t)$$
$$\pi_{2,3} : C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^3 \to C_T^{\text{Imm}}(\mathbb{R}, V) \times \mathbb{R}^2,\ (\gamma, r, s, t) \mapsto (\gamma, s, t)$$

These are continuous maps. Then NotCrossPair$_V$ is open by proposition A.3.3, and as noted before the proposition we have

$$\text{NotCrossTriple}_V =$$
$$\pi_{1,2}^{-1}(\text{NotCrossPair}_V) \cup \pi_{1,3}^{-1}(\text{NotCrossPair}_V) \cup \pi_{2,3}^{-1}(\text{NotCrossPair}_V),$$

which is visibly open. □

Proposition 3.14. *Let*

$$\text{AtMostDouble}_V =$$

$$\{(\gamma, s, t) \in C^{\text{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}^2 \mid (s,t) \textit{ is at most a double crossing of } \gamma\}.$$

Then AtMostDouble_V *is an open subset of* $C^{\text{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}$.

Proof. This can be proved from proposition A.3.13 in much the same way as proposition A.3.4 is deduced from proposition A.3.3. □

Proposition 3.15. *Let* $\gamma \in C^{\text{Imm}}_T(\mathbb{R}, V)$ *and suppose that* (s,t) *is a crossing pair of* γ *which is transversal and at most a double crossing. Then there is a neighbourhood* U *of* γ *and a neighbourhood* W *of* (s,t) *such that if* $\beta \in U$ *then* β *has at most one crossing pair on* W*, and any crossing pair* $(s',t') \in W$ *of* β *on* W *is a transversal double crossing.*

Proof. By proposition A.3.14 we can let U' be a neighbourhood of γ and W' be a neighbourhood of (s,t) such that if $\beta \in U'$ and $(s',t') \in W'$ then (s',t') is at most a double crossing of β. We have that $\gamma'(s)$ and $\gamma'(t)$ are non zero and non parallel so by proposition A.3.9 we can find $\delta > 0$ such that if we have curves $\alpha : I \to \mathbb{R}^3 n$, $\beta : J \to \mathbb{R}^3$ with I and J proper intervals with $\|\alpha'(x) - u\| < \delta$ for $x \in I$ and $\|\beta'(y) - v\| < \delta$ for $y \in J$, then there is at most one pair $(x,y) \in I \times J$ with $\alpha(x) = \beta(y)$, and if there is such a pair then $\alpha'(s)$ and $\beta'(t)$ are non zero and non parallel.

Pick an open interval W_s around s such that if $x \in W_s$ then $\|\gamma'(x) - \gamma'(s)\| < \frac{\delta}{2}$, and similarly pick an open interval W_t around t such that if $y \in W_t$ then $\|\gamma'(y) - \gamma'(t)\| < \frac{\delta}{2}$. We can assume that $W_s \times W_t \subseteq W'$. Then if $\beta \in C^{\text{Imm}}_T(\mathbb{R}, V)$ with $\|\beta - \gamma\|_{C^1} < \frac{\delta}{2}$, then for $x \in W_s$ we have $\|\beta'(x) - \gamma'(s)\| < \frac{\delta}{2} + \|\gamma'(x) - \gamma'(s)\| < \delta$, and similarly for $y \in W_t$ we have $\|\beta'(y) - \gamma'(t)\| < \delta$. Thus β has at most one crossing pair no $W_s \times W_t$, which is transversal by the choice of δ, and a double crossing by the choice of W'. □

Proposition 3.16. *For $\gamma \in C^{\mathrm{Imm}}_T(\mathbb{R}, V)$, say that $(s,t) \in \mathbb{R}^2$ is* unproblematic *for γ if either (s,t) is not a crossing pair of γ, or it is a transversal double crossing pair. Then if we let*

$$\mathrm{Unproblem}_V = \{(\gamma, s, t) \in C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}^2 \mid (s,t) \textit{ is unproblematic for } \gamma\},$$

then $\mathrm{Unproblem}_V$ *is an open subset of $C^{\mathrm{Imm}}_T(\mathbb{R}, V) \times \mathbb{R}^2$.*

Proof. We need to show that if $(\gamma, s, t) \in \mathrm{Unproblem}_V$ then there is an open neighbourhood W of (γ, s, t) such that $W \subseteq \mathrm{Unproblem}_V$. But for (s,t) which is not a crossing pair of γ this is proved in proposition A.3.3, and for (s,t) which is a transversal double crossing this is proved in proposition A.3.15. □

Theorem 3.17. *The set of $\gamma \in C^{\mathrm{Imm}}_T(\mathbb{R}, V)$ which have only double and transversal crossings is open.*

Proof. This follows from proposition A.3.16 in the same way that theorem A.3.6 follows from proposition A.3.3. □

Proposition 3.18. *Let P be a plane in $\mathbb{R}^3$. Then for all $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{R}^3)$ we have $\|\beta_P - \gamma_P\|_{C^1} \leqslant \|\beta - \gamma\|_{C^1}$. In particular, the function $\gamma \mapsto \gamma_P$ is continuous.*

Proof. We define the tangent space T_P of P to be $\{u - v \mid u, v \in P\}$, which is a plane parallel to P through 0. We let π_{T_P} be the projection onto T_P. For any smooth curve γ we have $\gamma'_P(t) = (\pi_P \circ \gamma)'(t) = \pi_{T_P}(\gamma'(t))$.

Then for all t we have

$$\|\beta_P(t) - \gamma_P(t)\| = \|\pi_P(\beta(t)) - \pi_P(\gamma(t))\| = \|\pi_P((\beta - \gamma)(t))\| \leqslant \|(\beta - \gamma)(t)\|$$

and similarly

$$\|\beta'_P(t) - \gamma'_P(t)\| = \|\pi_{T_P}(\beta'(t)) - \pi_{T_P}(\gamma'(t))\| = \|\pi_{T_P}((\beta - \gamma)'(t))\| \leqslant \|(\beta - \gamma)'(t)\|.$$

Thus

$$\|(\beta_P - \gamma_P)(t)\| + \|(\beta_P(t) - \gamma_P)'(t)\| \leqslant \|(\beta - \gamma)(t)\| + \|(\beta - \gamma)'(t)\| \leqslant \|\beta - \gamma\|_{C^1}$$

which proves the result. □

Proposition 3.19. *Let P be a plane, p_0 a point in P. The set of smooth curves $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{R}^3)$ such that for all t, $\gamma_P(t) \neq p_0$ is open and dense.*

Proof. First, openness. Suppose we have $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{R}^3)$ such that for all t $\gamma_P(t) \neq p_0$. The function $t \mapsto \|\gamma_P(t) - p_0\|$ is thus continuous and periodic on $\mathbb{R}$, and everywhere positive, so there is $\delta > 0$ such that $\|\gamma_P(t) - p_0\| \geqslant \delta$ for all t. Then if $\beta \in C^\infty_T(\mathbb{R}, \mathbb{R}^3)$ with $\|\beta - \gamma\|_{C^1} < \delta$ then for all t we have $\|\beta_P(t) - \gamma_P(t)\| \leqslant \|\beta - \gamma\|_{C^1}$ by proposition A.3.18 so

$$\|\beta_P(t) - p_0\| \geqslant \|\gamma_P(t) - p_0\| - \|\beta_P(t) - \gamma_P(t)\| \geqslant \delta - \|\beta - \gamma\|_{C^1} > 0$$

so that we have $\beta_P(t) \neq p_0$ for all t, as required.

Now for denseness. Let $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{R}^3)$. Then $\text{Image}(\gamma_P)$ cannot contain any ball around p_0, by proposition A.2.9. Thus given $\epsilon > 0$, we can find $q \in P \setminus \text{Image}(\gamma_P)$ with $\|q - p_0\| < \epsilon$. Then $\gamma + (p_0 - q)$ is a smooth curve in $C^\infty_T(\mathbb{R}, \mathbb{R}^3)$, with $\|(\gamma + (p_0 - q)) - \gamma\|_{C^1} < \epsilon$, and for all t, $(\gamma + (p_0 - q))_P(t) = \gamma_P(t) + p_0 - q \neq p_0$ since then we would have $\gamma_P(t) = q$. □

Proposition 3.20. *The set $C^{\text{Imm}}_T(\mathbb{R}, \mathbb{R}^2)$ is dense in $C^\infty_T(\mathbb{R}, C^\infty_T(\mathbb{R}, \mathbb{R}^3))$.*

Proof. Let $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{C})$ and let $\epsilon > 0$. The function

$$\alpha : t \mapsto -\frac{T}{2\pi i}\gamma'(t) \cdot e^{\frac{-2\pi i t}{T}}$$

is a smooth curve so by proposition A.2.9 its image contains no open ball in $\mathbb{C}$ around

0. Let $z \in \mathbb{C}$ be such that $\|z\|(1 + |\frac{2\pi i}{T}|) < \epsilon$, with $z \notin \text{Image}(\alpha)$. Let

$$\beta(t) = ze^{\frac{2\pi it}{T}},$$

so $\beta \in C^\infty_T(\mathbb{R}, \mathbb{C})$,

$$\beta'(t) = \frac{2\pi i}{T} ze^{\frac{2\pi it}{T}}.$$

Thus $\|\beta\|_{C^1} \leqslant \|z\| + \|\frac{2\pi i}{T} z\| < \epsilon$. Thus $\|(\gamma + \beta) - \gamma\|_{C^1} < \epsilon$. Moreover for all $t \in \mathbb{R}$ we have

$$\begin{gathered} z \neq -\frac{T}{2\pi i}\gamma'(t) \cdot e^{\frac{-2\pi it}{T}} \\ \text{so } \frac{2\pi i}{T} ze^{\frac{2\pi it}{T}} \neq -\gamma'(t) \\ \text{so } (\beta + \gamma)'(t) \neq 0. \end{gathered}$$

Thus $\beta + \gamma$ is an element of $C^{\text{Imm}}_T(\mathbb{R}, \mathbb{R}^2)$ within ϵ of γ. That proves the claim. □

Proposition 3.21. *Let $\gamma \in C^{\text{Imm}}_T(\mathbb{R}, \mathbb{R}^2)$ such that γ has only double and transversal crossing points. Let $c < d < c + T$ such that $\gamma(c) \neq \gamma(d)$. Then there are $\delta, \epsilon > 0$ such that if $0 < h < \epsilon$ and $\beta : [c - h, d + h] \to \mathbb{R}^2$ is smooth with $\beta|_{[c,d]} = \gamma|_{[c,d]}$ and $\|\beta'(t) - \gamma'(c)\| < \delta$ for $t \in [c - h, c]$ and $\|\beta'(t) - \gamma'(d)\| < \delta$ for $t \in [d, d + h]$ then β is a smooth immersion and has no crossing points on $[c - h, c) \cup (d, d + h]$.*

Proof. We pick $\delta_c > 0$ small enough that if $\|u - \gamma'(c)\| < \delta_c$ then $u \cdot \gamma'(c) > 0$, and pick $\epsilon_c > 0$ small enough that if $|x - c| < \epsilon_c$ then $|\gamma'(x) - \gamma'(c)| < \delta_c$, and $\gamma(d) \notin \gamma((c - \epsilon_c, c + \epsilon_c))$.

If c is a crossing point of $\gamma|_{[c,d]}$ then we add extra conditions on the choice of δ_c and ϵ_c. We have $\gamma(c) = \gamma(s_c)$ for some $s_c \in (c, d)$, and $\gamma'(c)$ and $\gamma'(s_c)$ are non parallel. Thus by proposition A.3.9 we can pick δ_c small enough that if we have curves $\rho : I \to \mathbb{R}^2$, $\sigma : J \to \mathbb{R}^2$ with I and J proper intervals with $\|\rho'(x) - \gamma'(c)\| < \delta_c$ for $x \in I$ and

$\|\sigma'(y) - \gamma'(s_c)\| < \delta_c$ for $y \in J$, then there is at most one pair $(x, y) \in I \times J$ with $\rho(x) = \sigma(y)$. Then we can pick ϵ_c small enough that if $|x-c| < \epsilon_c$ then $\|\gamma'(x)-\gamma'(c)\| < \delta_c$ and if $|x - s_c| < \epsilon_c$ then $\|\gamma'(x) - \gamma'(s_c)\| < \delta_c$. WLOG we may assume that ϵ_c is small enough that $(s_c - \epsilon_c, s_c + \epsilon_c) \subseteq (c, d)$.

Similarly we pick $\delta_d > 0$ small enough that if $\|u - \gamma'(d)\| < \delta_d$ then $u \cdot \gamma'(d) > 0$, and pick $\epsilon_d > 0$ small enough that if $|x - d| < \epsilon_d$ then $|\gamma'(x) - \gamma'(d)| < \delta_d$, and $\gamma(c) \notin \gamma((d - \epsilon_d, d + \epsilon_d))$.

If d is a crossing point of $\gamma|_{[c,d]}$ then we add extra conditions on the choice of δ_d and ϵ_d. We have $\gamma(d) = \gamma(s_d)$ for some $s_d \in (c, d)$, and $\gamma'(d)$ and $\gamma'(s_d)$ are non parallel. Thus by proposition A.3.9 we can pick δ_d small enough that if we have curves $\rho : I \to \mathbb{R}^2$, $\sigma : J \to \mathbb{R}^2$ with I and J proper intervals with $\|\rho'(x) - \gamma'(d)\| < \delta_d$ for $x \in I$ and $\|\sigma'(y) - \gamma'(s_d)\| < \delta_d$ for $y \in J$, then there is at most one pair $(x, y) \in I \times J$ with $\rho(x) = \sigma(y)$. Then we can pick ϵ_d small enough that if $|x - d| < \epsilon_d$ then $\|\gamma'(x) - \gamma'(d)\| < \delta_c$ and if $|x - s_d| < \epsilon_d$ then $\|\gamma'(x) - \gamma'(s_d)\| < \delta_d$. WLOG we may assume that ϵ_d is small enough that $(s_d - \epsilon_d, s_d + \epsilon_d) \subseteq (c, d)$.

If c is a crossing point of $\gamma|_{[c,d]}$, take $U_c = ((s_c - \epsilon_c, s_c + \epsilon_c) \cup (c - \epsilon_c, c + \epsilon_c))$, otherwise take $U_c = (c - \epsilon_c, c + \epsilon_c)$. Similarly for U_d. Thus U_c and U_d are open and $\gamma(c) \notin \gamma([c, d] \setminus U_c)$, and $\gamma(d) \notin \gamma([c, d] \setminus U_d)$, so since $\gamma([c, d] \setminus U_c)$ and $\gamma([c, d] \setminus U_d)$ are compact we can find $\eta > 0$ such that $B_{2\eta}(\gamma(c))$ is disjoint from $\gamma([c, d]\setminus U_c)$ and $B_{2\eta}(\gamma(d))$ is disjoint from $\gamma([c, d] \setminus U_d)$. Now take $\epsilon \leqslant \min(\epsilon_c, \epsilon_d)$ such that $\epsilon(\|\gamma'(c)\| + \delta_c) < \eta$ and $\epsilon(\|\gamma'(d)\| + \delta_d) < \eta$. Take $\delta = \min(\delta_c, \delta_d)$.

Now suppose $0 < h < \epsilon$ and $\beta : [c - h, d + h] \to \mathbb{R}$ is smooth with $\beta|_{[c,d]} = \gamma|_{[c,d]}$ and $\|\beta'(t) - \gamma'(c)\| < \delta$ for $t \in [c - h, c]$ and $\|\beta'(t) - \gamma'(d)\| < \delta$ for $t \in [d, d + h]$. First note that if $x \in [c - h, c]$ then by assumption $\|\beta'(x) - \gamma'(c)\| < \delta$, and if $x \in [c, c + \epsilon_c)$ then $x \in [c, d]$ so $\|\beta'(x) - \gamma'(c)\| = \|\gamma'(x) - \gamma'(c)\| < \delta_c$. Thus if $x \in [c - h, c + \epsilon_c)$ then $\|\beta'(x) - \gamma'(c)\| < \delta_c$. Thus if $x \in [c - h, c + \epsilon_c)$ then $\beta'(x) \cdot \gamma'(c) > 0$. Similarly if $y \in (d - \epsilon_d, d + h]$ then $\|\beta'(y) - \gamma'(d)\| < \delta_d$ and $\beta'(y) \cdot \gamma'(d) > 0$.

This means that β is a smooth immersion, since if $x \in [c-h, c]$ then $\beta'(x) \cdot \gamma'(c) > 0$ so $\beta'(x) \neq 0$, if $y \in [d, d+h]$ then $\beta'(y) \cdot \gamma'(d) > 0$ so $\beta'(y) \neq 0$, and if $x \in [c, d]$ then $\beta'(x) = \gamma'(x) \neq 0$.

Now we show that β has no crossing points on $[c-h, c) \cup (d, d+h]$. We have that if $x \in [c-h, c+\epsilon_c)$ then $\beta'(x) \cdot \gamma'(c) > 0$, so by proposition A.2.3 $\beta|_{[c-h,c+\epsilon_c)}$ is injective. Thus β has no crossing pair (x, y) with $x \in [c-h, h)$, $y \in (c-\epsilon, c+\epsilon)$. Thus if c is not a crossing point of γ then β has no crossing pairs (x, y) with $x \in [c-h, c)$, $y \in U_c$. Suppose on the other hand that c is a crossing point of $\gamma|_{[c,d]}$, with s_c as above. Then $\beta(c) = \beta(s_c)$. We have that if $x \in [c-h, c+\epsilon_c)$ then $\|\beta'(x) - \gamma'(c)\| < \delta_c$, and if $y \in (s_c - \epsilon_c, s_c + \epsilon_c)$ then $\|\gamma'(y) - \gamma'(s_c)\| < \delta_c$. Thus by the choice of δ_c, there is at most one pair $(x, y) \in [c-h, c+\epsilon_c) \times (s_c - \epsilon_c, s_c + \epsilon_c)$ with $\beta(x) = \beta(y)$. Thus since $\beta(c) = \beta(s_c)$, we have that if $x \in [c-h, c)$ and $y \in (s_c - \epsilon_c, s_c + \epsilon_c)$ then $\beta(x) \neq \beta(y)$. Then since in this case $U_c = ((s_c - \epsilon_c, s_c + \epsilon_c) \cup (c - \epsilon_c, c + \epsilon_c))$, we again obtain that β has no crossing pairs (x, y) with $x \in [c-h, c)$ and $y \in U_c$; so whether or not c is a crossing point of $\gamma|_{[c,d]}$, β has no crossing pairs (x, y) with $x \in [c-h, c)$ and $y \in U_c$.

Next suppose $x \in [c-h, c]$ and $y \in [c, d] \setminus U_c$. Then we have $\|\gamma(y) - \gamma(c)\| \geqslant 2\eta$, and

$$\|\beta(x) - \gamma(c)\| \leqslant \sup_{t \in [c-h,c]} \|\beta'(t)\||t - c| \leqslant (\|\beta'(c)\| + \delta_c)\epsilon < \eta$$

by the choice of ϵ. Thus $\|\beta(y) - \beta(x)\| = \|\gamma(y) - \beta(x)\| > \eta$. In particular $\beta(y) \neq \beta(x)$. Also taking $y = d$ we have that $\|\beta(d) - \beta(x)\| > \eta$, and by a similar argument we obtain that if $y \in [d, d+h]$ then $\|\beta(y) - \beta(d)\| < \eta$. Thus if $x \in [c-h, c)$ and $y \in [d, d+h]$ then $\beta(x) \neq \beta(y)$. Thus for all $x \in [c-h, c)$ and $y \in [c-h, d+h]$ we have that if $x \neq y$ then $\beta(x) \neq \beta(y)$. Thus $\beta_{[c-h,d+h]}$ has no crossing points on $[c-h, c)$. The proof that it has no crossing points on $(d, d+h]$ is similar. □

Proposition 3.22. *Let $\gamma \in C^\infty_T(\mathbb{R}, \mathbb{R}^2)$. Let $a < b < a + T$ and suppose $\beta : U \to \mathbb{R}^2$ is smooth with U an open interval containing $[b-T, a]$, and such that $\beta|_{[b-T,a]} = \gamma|_{[b-T,a]}$.*

Let $\epsilon > 0$. Then there is $\delta > 0$ such that for all $h \in (0, \delta)$ there is $\alpha \in C^\infty_T(\mathbb{R}, \mathbb{R}^2)$ such that $\alpha|_{[b-T-h,a+h]} = \beta|_{[b-T-h,a+h]}$, $\alpha|_{[a+2h,b-2h]} = \gamma|_{[a+2h,b-2h]}$ and $\|\alpha - \gamma\|_{C^1} < \epsilon$.

Proof. Let γ, a, b, U, β be as described. Let $\eta > 0$ such that $[b - T - \eta, a + \eta] \subseteq U$, and $a + \eta < b - \eta + T$.

We can find a family $\phi_h : (b - T - \eta, a + \eta) \to \mathbb{R}$ of smooth functions for $h \in (0, \frac{\eta}{2})$ such that $0 \leqslant \phi_h(t) \leqslant 1$ for all t, $\phi_h(t) = 0$ for $t \leqslant b - T - 2h$ or $t \geqslant a + 2h$, $\phi_h(t) = 1$ for $b - T - h \leqslant t \leqslant a + h$, and with a constant K such that $|\phi_h'(t)| \leqslant \frac{K}{h}$ for all h and t. Then for each h the map

$$[a + \eta - T, b - \eta] \to \mathbb{R},\ t \mapsto \begin{cases} \gamma(t) + \phi_h(t)(\beta(t) - \gamma(t)) & \text{if } t \in (b - T - \eta, a + \eta) \\ \gamma(t) & \text{if } t \notin [b - T - 2h, a + 2h] \end{cases}$$

is well defined and smooth, and extends to a smooth T-periodic map β_h. If $t \in [a + \eta - T, b - \eta]$ we have

$$\beta_h'(t) = \begin{cases} \gamma'(t) + \phi_h'(t)(\beta(t) - \gamma(t)) + \phi_h(t)(\beta'(t) - \gamma'(t)) & \text{if } t \in (b - T - \eta, a + \eta) \\ \gamma'(t) & \text{if } t \notin [b - T - 2h, a + 2h]. \end{cases}$$

We have $(\beta - \gamma)''(a) = (\beta - \gamma)''(b) = 0$ so there is $\rho > 0$ such that if $h \in (0, \rho)$ then $|\beta'(a+h) - \gamma'(a+h)| < h$, $|\beta'(b-T-h) - \gamma'(b-T-h)| < h$, $|\beta(a+h) - \gamma(a+h)| < h^2$ and $|\beta(b - T - h) - \gamma(b - T - h)| < h^2$. We have $\beta_h(t) = \gamma(t)$ for $t \in [b - T, a]$ and $t \in [a + 2h, b - 2h]$, so letting $0 < h < \frac{\rho}{2}$ and $A = [a, a + 2h] \cup [b - T - 2h, c]$ we have for all t that $|\beta_h(t) - \gamma(t)| \leqslant \sup_A \|\beta(t) - \gamma(t)\| \leqslant 4h^2$ and

$$\begin{aligned} |\beta_h'(t) - \gamma'(t)| &= |\phi_h'(t)(\beta(t) - \gamma(t)) + \phi_h(t)(\beta'(t) - \gamma'(t))| \\ &\leqslant \frac{K}{h} \sup_A \|\beta(t) - \gamma(t)\| + \sup_A \|\beta'(t) - \gamma'(t)\| \\ &\leqslant \frac{K}{h} \cdot 4h^2 + 2h \end{aligned}$$

so we obtain $\|\beta_h - \gamma\|_{C^1} \leqslant 4h^2 + 4Kh + 2h$ which tends to zero as h tends to zero. Thus given $\epsilon > 0$ we can find $\delta \in (0, \frac{\eta}{2})$ such that if $h \in (0, \delta)$ then we have $\|\beta_h - \gamma\|_{C^1} < \epsilon$, and this satisfies $\beta_h|_{[b-T-h,a+h]} = \beta_{[b-T-h,a+h]}$ and $\beta_h|_{[a+2h,b-2h]} = \gamma|_{[a+2h,b-2h]}$. □

Proposition 3.23. *Let $\gamma : I \to \mathbb{R}^2$ be an injective smooth immersion with I a compact interval, and let $\sigma_j : V_j \to \mathbb{R}^2$ be a smooth immersion for each $j \in J$ with each V_j an open interval and J a countable set. Let $X \subseteq \mathbb{R}^2$ be countable. Let $a < b \in I$ and let $\delta, \epsilon > 0$ with $\delta < \frac{b-a}{2}$. Then there is an injective smooth immersion $\beta : I \to \mathbb{R}^2$ with $\beta(t) = \gamma(t)$ for $t \notin (a, b)$, $\|\beta - \gamma\|_{C^1} < \epsilon$, $\beta'(t) = \gamma'(t)$ for $t \in [a + \delta, b - \delta]$, and such that for all $t \in [a + \delta, b - \delta]$ we have $\beta(t) \notin X$, and if $\beta(t) = \sigma_j(s)$ for any j then $\beta'(t)$ and $\sigma_j'(s)$ are not parallel.*

Proof. Here we write $v \parallel u$ to signify that the vectors v and u are parallel (or at least one is zero).

By proposition A.3.7 we can find $\eta > 0$ such that if $\beta : I \to \mathbb{R}^2$ is smooth with $\|\beta - \gamma\|_{C^1} < \eta$ then β is an injective smooth immersion.

Let $h : \mathbb{R} \to \mathbb{R}$ be a smooth function with $0 \leqslant h(t) \leqslant 1$ for all t, $h(t) = 0$ for $t \leqslant a$ and $t \geqslant b$ and $h(t) = 1$ for $t \in [a + \delta, b - \delta]$. If $z \in \mathbb{R}^2$ let

$$\gamma_z : I \to \mathbb{R}^2,\ t \mapsto \gamma(t) + h(t)z.$$

We have $\gamma_z(t) = \gamma(t)$ for $t \notin (a, b)$ and $\gamma_z'(t) = \gamma'(t)$ for $t \in [a + \delta, b - \delta]$. We also have $\|\gamma_z(t) - \gamma(t)\| \leqslant \|z\|$ for all t, and $\|\gamma_z'(t) - \gamma'(t)\| = \|h'(t)\|\|z\|$. Thus if we pick K with $K \geqslant h'(t)$ for all t then we have $\|\gamma_z - \gamma\|_{C^1} \leqslant (1 + K)\|z\|$, so if $\|z\| < \frac{\max(\eta,\epsilon)}{1+K}$ then γ_z is an injective smooth immersion and $\|\gamma_z - \gamma\|_{C^1} < \epsilon$. Thus letting $\eta' = \frac{\max(\eta,\epsilon)}{1+K}$ we just need to find z with $\|z\| < \eta'$ such that if $t \in [a + \delta, b - \delta]$ then $\gamma_z(t) \notin X$, and if $\gamma_z(t) = \sigma_j(s)$ for some j then $\gamma_z'(t)$ and $\sigma_j'(t)$ are not parallel.

For each $x \in X$ the image of the curve $(a, b) \to \mathbb{R}^2,\ t \mapsto x - \gamma(t)$ has measure zero by proposition A.2.9, so $\{x - \gamma(t) \mid t \in [a + \delta, b - \delta],\ x \in X\}$ has measure zero.

Now for each $j \in J$, consider the map $R_j : V_j \times (a,b) \to \mathbb{R}^2$, $(s,t) \mapsto \sigma_j(s) - \gamma(t)$, which is smooth. We have $\frac{\partial R_j}{\partial t}|_{s,t} = -\gamma'(t)$, $\frac{\partial R_j}{\partial s}|_{s,t} = \sigma_j'(s)$. Thus if $\sigma_j'(s) \parallel \gamma'(t)$ then the Jacobian of R_j at (s,t) has rank $\leqslant 1$, so (s,t) is a critical point of R_j, so $R_j(s,t)$ is a critical value of R_j. Thus by Sard's theorem $\{R_j(s,t) \mid \sigma_j'(s) \parallel \gamma'(t)\}$ has measure zero, so

$$\{z \mid \exists j \in J,\, t \in [a+\delta, b-\delta],\, s \in V_j \text{ with } z = \sigma_j(s) - \gamma(t),\, \sigma_j'(s) \parallel \gamma'(t)\}$$

has measure zero.

Thus the union of the above two sets has measure zero, so we can find z with $\|z\| < \eta'$ and such that for all $t \in [a+\delta, b-\delta]$, for all $x \in X$, $x - \gamma(t) \neq z$, and for all such t, all $j \in J$ and all $s \in V_j$, if $z = \gamma(t) - \sigma_j(s)$ then $\gamma'(t) \nparallel \sigma_j'(s)$. But the former just says that if $t \in [a+\delta, b-\delta]$ and $x \in X$ then we have $\gamma_z(t) = \gamma(t) + z \neq x$, so $\gamma_z(t) \notin X$. And the latter just says that if $t \in [a+\delta, b-\delta]$ and $j \in J$ and $s \in V_j$ with $\gamma_z(t) = \sigma_j(s)$ then $\gamma_z'(t) = \gamma'(t) \nparallel \sigma_j'(t)$. Thus taking β to be γ_z, we are done. □

For the next propositions, we extend the notions of transverse crossing and double crossing to curves with compact domain. Suppose that $\gamma : A \to \mathbb{R}^n$ is a C^1 immersion with A a compact proper interval. If (s,t) is a crossing pair of γ, we say that (s,t) is a **transverse** crossing if $\gamma'(s)$ and $\gamma'(t)$ are non parallel. We say that (s,t) is a **double** crossing if when we have $r \in A$ with $\gamma(r) = \gamma(s) = \gamma(t)$ then $r = s$ or $r = t$.

We also introduce the notion of a union-minimal family of sets. If $(U_i)_{i\in I}$ is a family of subsets of X, we say that $(U_i)_{i\in I}$ is **union-minimal** if for all proper subsets J of I we have $\bigcup_{i\in J} U_i \subsetneq \bigcup_{i\in I} U_i$. If $(U_i)_{i\in I}$ is a finite family of open intervals in $\mathbb{R}$ which is union-minimal then it is easy to see that for all i we have $U_i \neq \varnothing$, we have that the values $(\inf(U_i))_{i\in I}$ are distinct, and that if we order the U_i as $U_0 \dots U_n$ with $\inf(U_m) < \inf(U_p)$ for $m < p$ then we have for $m = 1 \dots (n-1)$ that $\inf(U_m) < \sup(U_{m-1}) < \inf(U_{m+1}) < \sup(U_m)$, and that $\inf(U_0) < \inf(U_1)$, $\sup(U_{n-1}) < \sup(U_n)$. Conversely if we have a

family $(U_m)_{m=0}^n$ of nonempty open intervals in $\mathbb{R}$ satisfying these conditions then it is easy to see that $(U_m)_{m=0}^n$ is union-minimal. If these conditions are satisfied then we call $(U_m)_{m=1}^n$ an ordered union-minimal family of open intervals.

Lemma 3.24. *Let $\gamma \in C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{R}^2)$. Let $\epsilon > 0$. Let (U, V, W) be an ordered union-minimal family of open intervals of length $< T$ such that:*

- $\gamma|_{\overline{U}}$ *has only transverse double crossings*
- $\gamma|_{\overline{V}}$ *is injective*
- $\gamma|_{\overline{W}}$ *is injective*
- $\inf(U) > \sup(V) - T$
- $\inf(V) > \sup(W) - T$

Then there are elements $\beta \in C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{R}^2)$, $a \in U \cap V$ and $b \in V \cap W$, and nonempty open intervals U' and W' such that:

- $\|\beta - \gamma\|_{C^1} < \epsilon$
- $\beta|_{[b-T,a]} = \gamma|_{[b-T,a]}$
- $\beta|_{\overline{U'}}$ *has only transverse double crossings*
- $\beta|_{\overline{W'}}$ *is injective*
- $U' \subseteq U \cup V$
- $W' \subseteq W$
- $U' \cup W' = U \cup V \cup W$
- $b \in W'$

Proof. By proposition A.3.11, $\gamma|_{\overline{U}}$ has finitely many crossing points, so we can pick $a \in U \cap V$ which is not a crossing point of $\gamma|_{\overline{U}}$. Since $\inf(V) > \sup(V) - T$ and $\inf(V) > \sup(W) - T$ we can choose a so that $a > \sup(V) - T$ and $a > \sup(W) - T$. We pick $b \in V \cap W$, and note that $b > \inf(W) > \sup(V) > a > \sup(V) - T > b - T$. Then we pick $\delta > 0$ such that $a + \delta \in U \cap V$, such that $\gamma|_{\overline{U}}$ has no crossing points on $[a, a + \delta]$, and such that $b - \delta \in V \cap W$. Then $\gamma([a, a + \delta])$ and $\gamma([\inf(U), \inf(V)])$ are disjoint, so we can find $\eta > 0$ such that $B_\eta(\gamma([a, a + \delta]))$ and $\gamma([\inf(U), \inf(V)])$ are disjoint, and $\eta < \epsilon$. Similarly we can find $\eta' > 0$ such that $B_{\eta'}(\gamma([b - \delta, b]))$ and $\gamma([\sup(V), \sup(W)])$ are disjoint. WLOG we assume that $\eta \leqslant \eta'$. Take

$$X = \{\gamma(s) \mid s \text{ is a crossing point of } \gamma|_{\overline{U}}\},$$

a finite set. Now since $\gamma|_{\overline{V}}$ is injective, by proposition A.3.23 we can find a smooth injective immersion $\rho : \overline{V} \to \mathbb{R}^2$ such that $\rho(t) = \gamma(t)$ for $t \in \overline{V} \setminus (a, b)$, such that $\|\rho - \gamma|_{\overline{V}}\|_{C^1} < \eta$, and such that if we have $t \in [a + \delta, b - \delta]$ then $\rho(t) \notin X$, and if $\rho(t) = \gamma(s)$ with $s \in \mathbb{R}$ then $\rho'(t)$ and $\gamma'(s)$ are not parallel.

Let β be the T-periodic smooth curve with $\beta|_{\overline{V}} = \rho$, $\beta|_{[b-T,a]} = \gamma|_{[b-T,a]}$. This is well defined since $\inf(V) > \sup(V) - T$ and $a > b - T$. Then β is a smooth immersion, with $\|\beta - \gamma\|_{C^1} < \epsilon$. Since $\inf(U) > \sup(V) - T$ we have $\inf(U) > b - T$ so that $\beta|_{[\inf(U),a]} = \gamma|_{[\inf(U),a]}$.

Suppose that (s, t) is a crossing pair of $\beta|_{[\inf(U),b-\delta]}$. We will argue that this crossing pair is transverse. Since $\beta|_{\overline{V}} = \rho$ is injective we cannot have both $s, t \in [\inf(V), b - \delta]$ and can assume WLOG that $s \in [\inf(U), \inf(V))$. Then since $\beta|_{[\inf(U),a]} = \gamma|_{[\inf(U),a]}$, if $t \in [\inf(U), a]$ then (s, t) is a crossing pair of $\gamma|_{\overline{U}}$ and so $\beta'(s) = \gamma'(s)$ and $\beta'(t) = \gamma'(t)$ are non parallel. Thus we may assume $t \in (a, b-\delta]$. But by the choice of η, $\gamma([\inf(U), \inf(V)])$ and $\rho([a, a + \delta])$ are disjoint, so since $\beta(s) = \beta(t)$ we must have $t \geqslant a + \delta$. But then by the choice of ρ we do indeed have that $\beta'(s) = \gamma'(s)$ and $\gamma'(t) = \rho'(t)$ are non parallel,

as required.

Next we show that if $r, s, t \in [\inf(U), b-\delta]$ are distinct then it's not the case that $\beta(r) = \beta(s) = \beta(t)$. Indeed, suppose that we had $r, s, t \in [\inf(U), b-\delta]$ with $\beta(r) = \beta(s) = \beta(t)$. Then since $\beta|_{[\inf(V), b-\delta]}$ is injective we have that at least 2 of r, s, t lie in $[\inf(U), \inf(V))$. WLOG $r, s \in [\inf(U), \inf(V))$, and so $\beta(r) = \beta(r) = \beta(s) = \gamma(s)$ and $\beta(r) \in X$. Then since γ has only double crossing points we cannot have $t \in [\inf(U), \inf(V)]$, and as above since $\beta(t) = \beta(r)$ cannot have $t \in [a, a+\delta]$, so that we must have $t \in [a+\delta, b-\delta$. But then $\rho(t) = \beta(t) = \beta(r) \in X$, contradicting the choice of ρ. Thus indeed if $r, s, t \in [\inf(U), b-\delta]$ are distinct then it's not the case that $\beta(r) = \beta(s) = \beta(t)$. Thus letting $c = b - \delta$ we have that $\beta|_{[\inf(A), c]}$ has only transverse double crossings.

Now we argue that $\beta|_{[b-\delta, \sup(W)]}$ is injective. So let $s < t \in [b-\delta, \sup(W)]$. Since $\sup(W) - T < a$ we have $\beta|_{[b, \sup(W)]} = \gamma|_{[b, \sup(W)]}$, so that $\beta|_{[b, \sup(W)]}$ is injective. Thus if $b \leqslant s < t$ then $\beta(s) \neq \beta(t)$ so we may assume $s < b$. Then since $\beta|_{[b-\delta, \sup(V)]} = \rho|_{[b-\delta, \sup(V)]}$ is injective we may assume $t > \sup(V)$. But then we have $s \in B_{\eta'}(\gamma([b-\delta, b]))$ and $t \in \gamma([\sup(V), \sup(W)])$ so that by the choice of η' we have $\beta(s) \neq \beta(t)$, and we are done. Thus $\beta|_{[b-\delta, \sup(W)]}$ is indeed injective.

Thus by proposition A.2.5 there is $w < b - \delta$ with $w \in V \cap W$ such that $\beta|_{[w, \sup(W)]}$ is injective. Then we let $U' = (\inf(U), b-\delta)$, $W' = (w, \sup(W))$, and then this choice of β, a, b, U' and W' satisfy the required conditions. □

Lemma 3.25. *Let $\gamma \in C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{R}^2)$. Let $\epsilon > 0$. Let U, V be nonempty open intervals of length $< T$ such that:*

- *$\gamma|_{\overline{U}}$ has only transverse double crossings*
- *$\gamma|_{\overline{V}}$ is injective*
- *$\sup(U) \in V$*

- $\sup(V) > \inf(U) + T$

Then there is an element $\beta \in C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{R}^2)$ *such that* $\|\beta - \gamma\|_{C^1} < \epsilon$ *and* β *has only transverse double crossings.*

Proof. The proof is very similar to lemma A.3.24. We have $\inf(V) + T > \sup(V) > \inf(U) + T$ so that $\inf(U) < \inf(V)$.

Since $\gamma|_{\overline{U}}$ has finitely many crossing points, so we can pick $a \in U \cap V$ which is not a crossing point of $\gamma|_{\overline{U}}$. Since $\inf(V) > \sup(V) - T$ we can choose a so that $a > \sup(V) - T$. Then we have $\inf(U) + T < \sup(V)$ so we can pick $b \in (\inf(U) + T, \sup(V))$ which is not a crossing point of $\gamma|_{\overline{U}+T}$, and we obtain $\inf(U) + T < b < \sup(V) < a + T$. Then we pick $\delta > 0$ such that $a + \delta \in U \cap V$, such that $\gamma|_{\overline{U}}$ has no crossing points on $[a, a + \delta]$, such that $b - \delta > \inf(U) + T$, and such that $\gamma|_{\overline{U}+T}$ has no crossing points on $[b, b - \delta]$. Then $\gamma([a, a + \delta])$, $\gamma([b - \delta - T, b - T])$, and $\gamma([\sup(V) - T, \inf(V)])$ are disjoint since $\gamma|_{\overline{U}}$ has no crossing points on $[a, a + \delta]$ or $[b - \delta - T, b - \delta]$. Thus we can find $\eta > 0$ such that $B_\eta(\gamma([a, a + \delta]))$, $B_\eta(\gamma([b - \delta - T, b - T]))$ and $\gamma([\sup(B) - T, \inf(V)])$ are disjoint, and $\eta < \epsilon$. Take

$$X = \{\gamma(s) \mid s \text{ is a crossing point of } \gamma|_{\overline{U}}\},$$

a finite set. Now since $\gamma|_{\overline{V}}$ is injective, by proposition A.3.23 we can find a smooth injective immersion $\rho : \overline{V} \to \mathbb{R}^2$ such that $\rho(t) = \gamma(t)$ for $t \in \overline{V} \setminus (a, b)$, such that $\|\rho - \gamma|_{\overline{V}}\|_{C^1} < \eta$, and such that if we have $t \in [a + \delta, b - \delta]$ then $\rho(t) \notin X$, and if $\rho(t) = \gamma(s)$ with $s \in \mathbb{R}$ then $\rho'(t)$ and $\gamma'(s)$ are not parallel.

Let β be the T-periodic smooth curve with $\beta|_{\overline{V}} = \rho$, $\beta|_{[b-T,a]} = \gamma|_{[b-T,a]}$. This is well defined since the length of V is $< T$ and $a > b - T$. Then β is a smooth immersion, with $\|\beta - \gamma\|_{C^1} < \epsilon$.

Suppose that (s, t) is a crossing pair of β. We will argue that this crossing pair is transverse. We can assume WLOG that $s, t \in (b - T, b]$ and that $s < t$. If $s, t \in [b - T, a]$

we are done since $\beta|_{[b-T,a]} = \gamma|_{[b-T,a]}$ and $\gamma|_{\overline{U}}$ has only transverse crossings. If $s, t \in \overline{V}$ then we are done since $\beta|_{\overline{V}} = \rho|_{\overline{V}}$ is injective. Thus we may assume $s < \inf(V)$, $t > a$. Then if $s \in (b - T, \sup(V) - T]$, $t \in [a, b]$ we have $\beta(s) = \rho(s + T) \neq \rho(t)$ since ρ is injective. Finally if $s \in [\sup(V) - T, \inf(V)]$ and $t \in [a, b]$ then by the choice of η we have $t \notin [a, a + \delta] \cup [b - \delta, b]$ so that $t \in [a + \delta, b - \delta]$. But then by the choice of ρ we have that $\beta'(s) = \gamma'(s)$ and $\beta'(t) = \rho'(t)$ are non parallel, as required.

Finally we show that if $r, s, t \in (b - T, b]$ are distinct then it's not the case that $\beta(r) = \beta(s) = \beta(t)$. Indeed, suppose that we had $r, s, t \in (b - T, b]$ distinct with $\beta(r) = \beta(s) = \beta(t)$. Then since $\beta|_{[\inf(V),b]}$ is injective we have that at least 2 of r, s, t lie in $(b - T, \inf(V)))$. WLOG $r, s \in (b - T, \inf(V))$. Thus $\gamma(r) = \beta(r) = \beta(s) = \gamma(s)$ and $\beta(r) \in X$. Then since γ has only double crossing points we cannot have $t \in (b - T, a]$, and by the choice of η cannot have $t \in (a, a + \delta) \cup (b - \delta, b]$ so must have $t \in [a + \delta, b - \delta]$. But then $\rho(t) = \beta(t) = \beta(r) \in X$, contradicting the choice of ρ. Thus indeed if $r, s, t \in (b - T, b]$ are distinct then it's not the case that $\beta(r) = \beta(s) = \beta(t)$.

Thus β has only transverse double crossings, as required. □

Theorem 3.26. *The set of $\beta \in C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{R}^2)$ which have only double and transversal crossings is dense.*

Proof. Let $\alpha \in C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{R}^2)$ and let $\epsilon > 0$. We seek $\beta \in C_T^{\mathrm{Imm}}(\mathbb{R}, \mathbb{R}^2)$ such that $\|\beta - \alpha\|_{C^1} < \epsilon$ and β has only double and transversal crossings. By proposition A.2.4 we can find a finite open cover $U_0, \ldots U_n$ of $[0, T]$ such that $\alpha|_{\overline{U_i}}$ is injective for each i. We may assume that the length of each U_i is $< T$, and that this cover is a minimal cover of $[0, T]$, and can also arrange things so that $n \geqslant 2$, for convenience. Since this is a minimal cover of $[0, T]$ it is certainly a union-minimal family. Then by reordering the U_i may assume that $i < j$ implies $\inf U_i < \inf U_j$, and thus that $(U_i)_{i=0}^n$ is an ordered union-minimal family. It can be seen that for each $i > 0$, $\inf(U_i) > 0$. By shrinking the U_i if necessary we may also assume that $\inf(U_0) > \sup(U_{n-1}) - T$, $\sup(U_n) < \inf(U_1) + T$.

We have that $\alpha|_{\overline{U_0}}$ is injective so has only transverse double crossings (since it has no crossing points), and $\alpha|_{\overline{U_1}}$, $\alpha|_{\overline{U_2}}$ are also injective. We also have $\inf(\overline{U_0}) > \sup(U_{n-1}) - T > \sup(\overline{U_1}) - T$ and $\inf(\overline{U_1}) > 0 > \sup(\overline{U_2}) - T$. Thus by lemma A.3.24 there is an element $\beta_1 \in C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^2)$ with $\|\beta_1 - \alpha\|_{C^1} < \frac{\epsilon}{n}$, and such that there are $a_1 \in U_0 \cap U_1$, $b_1 \in U_1 \cap U_2$ with b and nonempty open intervals I_1, W_1 such that:

- $\beta_1|_{[b_1-T,a_1]} = \gamma|_{[b_1-T,a]}$
- $\beta_1|_{\overline{I_1}}$ has only transverse double crossings
- $\beta_1|_{\overline{W_1}}$ is injective
- $I_1 \subseteq U_0 \cup U_1$
- $W_1 \subseteq U_2$
- $I_1 \cup W_1 = U_0 \cup U_1 \cup U_2$
- $b_1 \in W_1$

Suppose that $n \geqslant 3$. Then for $i = 3 \ldots n$, since $b_1 < \inf(U_i) < \sup(U_i) \leqslant \sup(U_n) < \inf(U_1) + T < a_1 + T$ we have that $\beta_1|_{\overline{U_i}} = \gamma|_{\overline{U_i}}$, and so $\beta_1|_{\overline{U_i}}$ is injective. Also $\inf(I_1) < \inf(W_1) < \sup(I_1) < \inf(U_3) < \sup(W_1) < \sup(U_3)$ and so (I_1, W_1, U_3) is an ordered union-minimal family. Finally $\inf(I_1) > \sup(U_2) - T = \sup(W_1) - T$, and $\inf(W_1) > \inf(U_2) > \inf(U_1) > \sup(U_n) - T \geqslant \sup(U_3) - T$. Thus we can use lemma A.3.24 again to obtain that there is an element $\beta_2 \in C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^2)$ with $\|\beta_2 - \beta_1\|_{C^1} < \frac{\epsilon}{n}$, and such that there are $a_2 \in I_1 \cap W_1$, $b_2 \in W_1 \cap U_3$ and nonempty open intervals I_2, W_2 such that:

- $\beta_2|_{[b_2-T,a_2]} = \beta_1|_{[b_2-T,a_2]}$
- $\beta_2|_{\overline{I_2}}$ has only transverse double crossings
- $\beta_2|_{\overline{W_2}}$ is injective
- $I_2 \subseteq I_1 \cup W_1 \subseteq U_0 \cup U_1 \cup U_2$

- $W_2 \subseteq U_3$
- $I_2 \cup W_2 = I_1 \cup W_1 \cup U_3 = U_0 \cup U_1 \cup U_2 \cup U_3$
- $b_2 \in W_2$

Thus $\|\beta_2 - \alpha\|_{C^1} < \frac{2\epsilon}{n}$, $a_2 \in U_1 \cap U_2$, $b_2 \in U_2 \cap U_3$.

Suppose that $n \geqslant 4$. Then for $i = 4 \dots n$, we have that $\beta_2|_{\overline{U_i}} = \beta_1|_{\overline{U_i}} = \gamma|_{\overline{U_i}}$ and so $\beta_2|\overline{U_i}$ is injective. We again have that (I_2, W_2, U_4) is an ordered union-minimal family, and that $\inf(I_2) > \sup(W_2) - T$, $\inf(W_2) > \sup(U_4) - T$. Thus again we can use lemma A.3.24 to obtain that there is an element $\beta_3 \in C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^2)$ with $\|\beta_3 - \beta_2\|_{C^1} < \frac{\epsilon}{n}$, and such that there are $a_3 \in I_2 \cap W_2$, $b_3 \in I_2 \cap U_4$ and nonempty open intervals I_3, W_3 such that

- $\beta_3|_{[b_3 - T, a_3]} = \beta_2|_{[b_3 - T, a_3]}$
- $\beta_3|_{\overline{I_3}}$ has only transverse double crossings
- $\beta_3|_{\overline{W_3}}$ is injective
- $I_3 \subseteq I_2 \cup W_2 \subseteq U_0 \cup U_1 \cup U_2 \cup U_3$
- $W_3 \subseteq U_4$
- $I_3 \cup W_3 = U_0 \cup U_1 \dots \cup U_4$
- $b_3 \in W_3$

Thus $\|\beta_3 - \alpha\|_{C^1} < \frac{3\epsilon}{n}$, $a_3 \in U_2 \cap U_3$, $b_3 \in U_3 \cap U_4$.

Continuing in this way we obtain ultimately an element $\beta_{n-1} \in C^{\mathrm{Imm}}_T(\mathbb{R}, \mathbb{R}^2)$ with $\|\beta_{n-1} - \gamma\|_{C^1} < \frac{(n-1)\epsilon}{n}$, such that there are $a_{n-1} \in U_{n-2} \cap U_{n-1}$, $b_{n-1} \in U_{n-1} \cap U_n$, and nonempty open intervals I_{n-1}, W_{n-1} such that

- $\beta_{n-1}|_{[b_{n-1} - T, a_{n-1}]} = \beta_{n-1}|_{[b_{n-1} - T, a_{n-1}]}$

- $\beta_{n-1}|_{\overline{I_{n-1}}}$ has only transverse double crossings
- $\beta_{n-1}|_{\overline{W_{n-1}}}$ is injective
- $I_{n-1} \subseteq U_0 \cup U_1 \cup \ldots \cup U_{n-1}$
- $W_{n-1} \subseteq U_n$
- $I_{n-1} \cup W_{n-1} = U_0 \cup U_1 \ldots \cup U_n$
- $b_{n-1} \in W_{n-1}$

But now we have $\sup(I_{n-1}) \in W_{n-1}$, $\sup(W_{n-1}) = \sup(U_n) > T > \inf(U_0) + T = \inf(I_{n-1}) + T$ so we can apply lemma A.3.25 and deduce that there is an element $\beta_n \in C_T^{\text{Imm}}(\mathbb{R}, \mathbb{R}^2)$ with $\|\beta_n - \beta_{n-1}\|_{C^1} < \frac{\epsilon}{n}$ and with β_n having only transverse and double crossings. But then $\|\beta_n - \alpha\|_{C^1} < \epsilon$ and we are done. □

Appendix B

Ontologically innocent second order logic

This appendix defends what I call restricted predicative second order logic as an ontologically innocent fragment of second order logic, which formalizes a conception of second order logic in which the second order quantifiers are open ended – interpretable at any point in time as ranging over the open formulae that have so far been formed, but not limited in their significance to any such restricted domain, and open to extension as the vocabulary expands. This is similar to the open ended interpretation of schematic logic that some authors have defended (Lavine 1998, §VII.4; Parsons 2007, §47), but resulting in a richer, more expressive fragment of full second order logic.

By an open formula we just mean a formula ϕ for which we have marked out a list $x_1, \ldots x_n$ of distinct variables as "arguments". We denote the result by $\phi(x_1, \ldots x_n)$. Then if $a_1 \ldots a_n$ are objects, $\phi(x_1, \ldots x_n)$ holds of $a_1 \ldots a_n$ iff ϕ is true when each x_i denotes a_i. If λ is a closed sentence, then we do not need to mark it with arguments – λ holds of $a_1 \ldots a_n$ iff ϕ is true.

The syntax of restricted predicative second order logic is the same as that of standard second order logic. One has a countably infinite stock of first order variables, denoted

$x, y, z, x_1, y', \ldots$, and for each natural number $n \geqslant 1$ a (disjoint) countably infinite stock of second order variables of arity n, commonly denoted $P, Q, R, P', Q_2, \ldots$. One specifies a language for the logic by specifying collections of (first order) constants, functions and relation symbols – each function and relation symbol having a specified arity $n \geqslant 1$. Terms are formed by applying function symbols to constants and first order variables as for first order logic. Then R is a relation symbol of arity n, and $t_1, \ldots t_n$ are terms, then $R(t_1, \ldots t_n)$ is an atomic sentence; similarly if P is a relation symbol of arity n, and $t_1, \ldots t_n$ are terms, then $P(t_1, \ldots t_n)$ is an atomic sentence. One obtains more complex sentences by joining sentences with propositional connectives, or affixing quantifiers, either first order $\forall x\ \forall y$ or second order $\forall P\ \forall Q$ of any arity.

If P and Q are second order variables of the same arity and P does not occur in ϕ in the scope of any quantifier that binds Q, then we write $\phi(Q|P)$ for the substitution of Q for P in ϕ, in the usual way.

Deductively, the difference between predicative second order logic and standard second order logic is in the comprehension scheme. In standard second order logic this scheme consists of the formulae

$$\exists P\ \forall x_1 \ldots x_n\ (P(x_1, \ldots x_n)\ \Leftrightarrow\ \phi)$$

for every n, every second order variable P of arity n, every n-tuple $x_1, \ldots x_n$ of distinct first order variables and every formula ϕ. In predicative second order logic this scheme is restricted only to those formulae ϕ which contain no second order quantifiers. In *restricted* predicative second order logic, this scheme is restricted only to those formulae ϕ which contain no second order variables at all, bound or free.

Apart from this change, the deductive rules for restricted predicative second order logic are the same as for standard second order logic. In the introduction and elimination rules for second order quantifiers, the second order variables of the appropriate arity play

the role of "terms".

The above was stated for a setting where we have a single type of first order variables, in addition to which we introduce second order variables. Sometimes we will also want to use restricted predicative second order logic in cases where there are multiple types of first order variables, or for instance a type of first order variables and then also plural variables. We call such initial types, to which we introduce in addition a type of second order variables, *object types*. In cases where there is more than one object type, the arity of a second order variable is not a number, but instead a list of object types, with atomic formulae obtained from second order variables by combining them with a list of terms of the appropriate types.

Importantly though no second order variables are allowed in the formula being used in an instance of comprehension in restricted predicative second order logic, free and bound variables of other kinds are: this formalizes the ability to talk for instance about the predicate "related to x", where x is a perhaps unknown or unnamed person. This applies even to variables that may have second order properties, as long as they are distinguished (of a different type) from the second order variables being used to represent open formulae. For instance if we accept plural logic, then we allow free (or bound) plural variables in instances of comprehension, representing the ability to form a predicate like "related to one of the xx's" where "xx" is a plural variable. The same goes if as well as our second order variables discussed here representing predicates or open formulae, we then introduce a further type of second order variables representing properties or relations as abstract entities independent of our ability to define them – as long as we take such entities to exist, and to properly be the subject of quantification, there is no problem with allowing them in instances of comprehension.

The interpretation of restricted predicative second order logic on which it is ontologically innocent is one where we make second order assertions that are valid when the domain of the second order quantifiers consists only of the open formulae we have so far

formed, but which will remain valid however this domain of quantification expands. Asserting a universally quantified second order statement thus demands that it will remain true as we state new formulae, and even expand our language to include new vocabulary.

To see how this works formally, if ϕ is a formula we can express, then we can state it in an instance of comprehension

$$\exists P \, \forall x_1 \ldots x_n \, (P(x_1, \ldots x_n) \; \Leftrightarrow \; \phi).$$

Then given a universally quantified formula $\forall Q \; \Psi(Q)$ where Q is of the same arity as P, we can reason about a P whose existence is asserted by this instance of comprehension, deducing $\Psi(P)$, and thus $\Psi(\phi)$ using the equivalence in the instance of comprehension.

We can give a simple semantics to model ontologically innocent interpretation of this logic. If ϕ is a formula and $x_1, \ldots x_n$ are distinct first order variables with $n \geqslant 1$, we call the pair $(\phi, (x_1, \ldots x_n))$ an *open formula*, and denote it by $\phi(x_1, \ldots x_n)$. We call n the arity of $\phi(x_1, \ldots x_n)$ Given a language $\mathcal{L}$ for the logic, by a structure we just mean the usual notion of first order structure for $\mathcal{L}$. By a *second order structure* we mean a pair (M, S) where M is a first order structure and S a set of open formulae which contains at least one open formula of each arity. In this set up, S can model the set of open formulae that we have formed at a given point in time.

The condition that S contain an open formula of each arity is important, as it means we can use the standard natural deduction rules for second order quantifiers instead of "inclusive" versions of those rules (inclusive being the term for a logic which allows for the possibility that the domain of quantification may be empty). However it may seem gratuitous to build this into our definition of second order structure, since it seems perfectly possible that we will *not* have managed at a given point in time to have formed an open formula of every arity; indeed this might seem more like a near certainty than a possibility, since forming infinitely many open formulae seems a hard task for us finite

beings. However as noted initially, any closed formula immediately delivers an open formula of every arity in a trivial way – where the resulting open formula (of each arity) does not actually depend on the values its arguments take. Thus to obtain a suitable S, we can take ourselves to be initially specifying some arbitrary closed formula λ, regarding it as providing an open formula of every arity. If one is uncomfortable with this move one could otherwise rephrase the deductive system as an inclusive logic (but that would complicate things).

Back to the semantics. By a *variable assignment* over such an (M, S) we mean a function v which assigns an element $v(x) \in M$ to each first order variable, and assigns to each second order variable P an open formula $v(P) \in S$ of the same arity.

We define a notion of satisfaction for a second order structure (M, S), a variable assignment v over it and a formula ϕ, in the usual way by induction on ϕ. We write this relation holding by (M, S), $v \vDash \phi$. The clauses for interpretation t^v of a first order term t with respect to variable assignment v are the usual ones, as is the clause for atomic formulae containing relation symbols. If P is a second order variable of arity n and $t_1 \ldots t_n$ are terms, then we define (M, S), $v \vDash P(t_1, \ldots t_n)$ to hold iff when we write $v(P)$ as $\phi(x_1, \ldots x_n)$, we have (M, S), $v(x_1 \mapsto t_1^v, \ldots x_n \mapsto t_n^v) \vDash \phi$. Then clauses for propositional connectives and first order quantifiers are as usual, and the clauses for the second order quantifiers are the obvious ones, for instance with (M, S), $v \vDash \forall P\, \phi$ iff for every $s \in S$ of the same arity as P, we have (M, S), $v(P \mapsto s) \vDash \phi$.

We define a sequent to be a pair (Γ, ϕ) where Γ is a finite set of formulae and ϕ a formula. We can write this as $\Gamma \vdash \phi$, or as $\gamma_1, \ldots \gamma_n \vdash \phi$ if $\Gamma = \{\gamma_1, \ldots \gamma_n\}$. If (M, S) is a structure and $\Gamma \vdash \phi$ a sequent, we say that (M, S) satisfies $\Gamma \vdash \phi$ if every variable assignment v over (M, S) which satisfies every element of Γ also satisfies ϕ.

Then the observation underlying the ontologically innocent status of restricted predicative second order logic is just that whenever we form a derivation D in the logic, the validity of the conclusion of D is ensured even when the domain of second order quanti-

fiers is restricted to those open formulae found in instances of predicative comprehension used in D – so that we never need to assume an infinity of formulae for our reasoning to be justified, and the formulae we have written down (or stated) always already suffice.

More formally, given an instance

$$\exists P\, \forall x_1 \ldots x_n\, (P(x_1, \ldots x_n) \;\Leftrightarrow\; \phi)$$

of restricted predicative comprehension, we define the corresponding open formula to be $\phi(x_1, \ldots x_n)$. We can take this open formula to be implicitly specified by this instance of comprehension – with ϕ being literally written out within it, and the list $x_1 \ldots x_n$ of variables also literally written out. Then given a derivation D in the logic, we define the set of open formulae S_D corresponding to D to contain open formulae corresponding to each instance of comprehension in D (and also to contain open formulae of the form $\lambda(y_1, \ldots y_n)$ of each arity for some distinguished, specified closed formula λ, as discussed above).

Then the argument is simply that for each M and D, if ϕ is the conclusion of D with live set of premises Γ, then (M, S_D) satisfies $\Gamma \vdash \phi$. We prove this by proving by induction on each line ψ of D that if Δ is the set of live premises at ψ, then (M, S_D) satisfies $\Delta \vdash \psi$. For lines that are premises this is trivial, and for lines deduced by any of the natural deduction rules this is just the standard case of soundness of those rules (though in a restricted domain, for the rules for second order quantification). The only potentially problematic case is that of lines which are instances of comprehension. So suppose that ψ is an instance of comprehension, more explicitly of the form

$$\exists P\, \forall x_1 \ldots x_n\, (P(x_1, \ldots x_n) \;\Leftrightarrow\; \phi).$$

By assumption the open formula $\phi(x_1, \ldots x_n)$ is in S_D. Suppose that v is any variable assignment over (M, S_D). We seek to argue that v satisfies ψ. Assigning value

$\phi(x_1, \ldots x_n) \in S_D$ to P, it suffices to show that for all $a_1 \ldots a_n \in M$, we have

$$(M, S_D),\, v(P \mapsto \phi(x_1, \ldots x_n), x_1 \mapsto a_1, \ldots x_n \mapsto a_n) \vDash P(x_1, \ldots x_n)$$
$$\Leftrightarrow (M, S_D),\, v(P \mapsto \phi(x_1, \ldots x_n), x_1 \mapsto a_1, \ldots x_n \mapsto a_n) \vDash \phi.$$

But this is trivial by the definition of satisfaction for $P(x_1, \ldots x_n)$, so we are done.

Thus we can use this logic to make statements and argue without being committed to any formulae beyond those which explicitly appear in instances of comprehension we appeal to. On the open ended conception of second order quantifiers, we can assert statements with universal second order quantifiers as long as those statements will remain true no matter how the vocabulary of our language expands, and what new open formulae we become able to form. This is entirely compatible with the above argument: though there was no explicit mention of the possibility of language change, one could have a derivation D with an initial segment in a restricted portion of the language, making universally quantified assertions which then have implications later in the derivation where a broader vocabulary is used.

As an example of the open ended use of the quantifiers, we might encounter a simply infinite sequence $\mathbb{N}$, with an initial object 0 and a successor operation given by S; then we could state that for any predicate P of elements of this sequence, if P applies to 0 and applies to $S(n)$ whenever it applies to n then it applies to all elements of the sequence:

$$\forall P\ (P(0) \wedge \forall n P(n) \rightarrow P(S(n))) \rightarrow \forall n P(n).$$

Asserting this requires being sure that it will remain true no matter how our vocabulary expands (so that the Sorites paradox poses a potential problem, unless we can somehow rule out predicates like "small" as illegitimate).

Though the above argument was given for a language with just two types of variables – first order and second order – one could give a similar argument if there were other

types, such as multiple different types of first order variables, or plural variables, or a different type of second order variable representing properties and relations as abstract entities. A little should be said about how restricted predicative second order logic interacts with other logics however.

A subtlety arises since a universally quantified statement $\forall R\ \phi[R]$ does not actually allow any substitution instance $\phi[\psi(x_1, \ldots x_n)]$ for any open formula $\psi(x_1, \ldots x_n)$ to be derived. Indeed one derives substitution instances via the comprehension scheme, which is restricted to instances without second order variables in restricted predicative second order logic, so that one can only directly derive $\phi[\psi(x_1, \ldots x_n)]$ from $\forall R\ \phi[R]$ when ψ does not contain any second order variables. Thus a universally quantified statement $\forall R\ \phi[R]$ does not actually have the usual force of an axiom scheme with template $\phi[R]$. This means that when we combine restricted predicative second order logic with another logic which uses an axiom scheme, such as plural logic, full second order logic, or double ancestral logic, there is a question about how the axiom scheme should be formalized: as a universally quantified statement of the form $\forall R\ \phi[R]$, or as the full infinite (and open ended) set of substitution instances of a template $\phi[R]$?

When giving a semantics for the joint logic (combining the above semantics for restricted predicative second order logic with the usual semantics for the other logic(s)), one can generally justify the full scheme rather than just the quantified formula. However there may be a worry that on the above semantics for restricted predicative second order logic, there is no fixed domain for the second order variables, so that a statement with bound second order variables does not have a fixed sense: thus for instance if $\chi(x)$ is an open formula containing bound predicate variables, it may be felt to be illegitimate to form the plurality of objects which satisfy $\chi(x)$, since which objects are amongst this plurality may shift as the domain of predicate variables expands. However we do not need to address this worry, since for the applications of joint logics in this chapter we only ever need free predicate variables – rather than bound predicate variables – in

the instances of the plural logic comprehension scheme and the double ancestral logic induction scheme, and such instances are legitimate, since for any well defined predicate R (no matter what the totality of such predicates is) if $\chi(x)$ is an open formula involving R then we can form the plurality of objects satisfying χ, or argue by induction along χ.

By considering a variant of the simply infinite sequence case, we can give an example which shows that the restriction in restricted predicative second order logic is genuine, and that restricted predicative second order logic is weaker than predicative second order logic. Suppose that we have not one simply infinite sequence but two, defined by predicates $\mathbb{N}_1, \mathbb{N}_2$, with initial elements $0_1, 0_2$ and successor operations S_1, S_2. Suppose we hope to define an isomorphism between $\mathbb{N}_1$ and $\mathbb{N}_2$. The standard way to do this is to go via an attempt, where an attempt is a binary relation R such that:

- R's domain is a subset of $\mathbb{N}_1$
- R's codomain is a subset of $\mathbb{N}_2$
- R is single valued on its domain
- if $S(n)$ is in R's domain then so is n
- R relates 0_1 to 0_2
- if R relates $n_1 \in \mathbb{N}_1$ to $n_2 \in \mathbb{N}_2$, and $S_1(n_1)$ is in R's domain, then R relates $S_1(n_1)$ to $S_2(n_2)$

Thus an attempt is a partial isomorphism, defined on an initial segment of $\mathbb{N}_1$. Then in predicative second order logic, we can prove that for every $n \in \mathbb{N}_1$ there is an attempt with n in its domain. The proof is by induction on n. For the induction step, we are given an attempt R with n in its domain, and seek to define an attempt R' with $S_1(n)$ in its domain. For this we can use predicative comprehension, taking R' to be the relation which relates all things related by R, and which also relates $S_1(n)$ to $S_2(m)$ where R

relates n to m. Key here is the ability to form an instance of comprehension in which a free predicate variable (R) appears – exactly what restricted predicative second order logic rules out. Having proved that for every $n \in \mathbb{N}_1$, an attempt exists which is defined at n, we can obtain an isomorphism $\mathbb{N}_1$ to $\mathbb{N}_2$ by quantifying over attempts – where the isomorphism relates n to m if there is an attempt which relates n to m.[1]

It is clear that this argument cannot go through in restricted predicative second order logic. Indeed we can define a second order structure whose base set is the disjoint union of two infinite sequences, thus giving an interpretation to $\mathbb{N}_1, \mathbb{N}_2, 0_1, 0_2, S_1, S_2$, and whose set of open formulae just consists of the open formulae of the form $\lambda(y_1, \ldots y_n)$ of each arity where λ is some closed formula. Then the statement that for all $n \in \mathbb{N}_1$ there is an attempt defined at n is trivially false in this second order structure.

[1] We can form an open formula defining this isomorphism in predicative second order logic, though we cannot prove the existence of it as the value of a second order variable, since that would require bound second order variables in the comprehension scheme.

Appendix C

Complete ordered field structure on a continuously ordered open interval

Here we suppose we are given the structure $R_<$, R_S of a continuously ordered open interval (as defined in section VI.4) on some domain, and will show how to define a field structure on this domain which, together with $R_<$, makes it into a complete ordered field.

We will write X for the domain of $R_<$, and write $<_X$ for $R_<$. We will write $\mathbb{N}$ for the domain of R_S, and write S for the successor function on $\mathbb{N}$ defined by R_S. Using double ancestral logic we can define addition and multiplication on $\mathbb{N}$ from S (this is the only use of double ancestral logic in the argument), which we write as $+_\mathbb{N}$ and $\times_\mathbb{N}$. We write $<_\mathbb{N}$ for the usual ordering on $\mathbb{N}$. We introduce the symbol α, which we use just to signify the inclusion $\mathbb{N} \hookrightarrow X$, but where we think of n and $\alpha(n)$ as playing very different roles: n plays the role of a natural number, and $\alpha(n)$ plays the role of an element of the totally ordered set X, with the totality of all $\alpha(n)$ being dense in X.

Since does not have endpoints, if $x \in X$ then there are $y, z \in X$ with $y <_X x$ and $z >_X x$. In particular since X is nonempty, it thus has at least two elements, so that its topology (the order topology on it) has base given by the open intervals of the form $(x, \infty) = \{y \in X \mid y >_X x\}$, $(\infty, x) = \{y \in X \mid y <_X x\}$ and $(x, y) = \{z \in X \mid x <_X z <_X y\}$. As initially noted, for any $x \in X$ there is $y \in X$ with $y >_X x$, so that $(x, \infty) \neq \varnothing$, and thus since the $\alpha(n)$ are dense in X there is $n \in \mathbb{N}$ such that $\alpha(n) \in (x, \infty)$. Similarly there is $n \in \mathbb{N}$ such that $\alpha(n) \in (\infty, x)$. Finally since X is densely ordered, for any $x, y \in X$ we have that $(x, y) \neq \varnothing$, so that there is $n \in \mathbb{N}$ with $\alpha(n) \in (x, y)$.

The usual argument that any nonempty dense, separable, complete totally ordered set X without endpoints is isomorphic to $(\mathbb{R}, <)$ works by setting up a correspondence between $\mathbb{Q}$ and the dense countable subset of X, and then extending this to all of $\mathbb{R}$ and X by completeness. The argument given here is basically similar, though we start with no copy of $\mathbb{R}$ available, and have to be more careful in places due to the limitations of the logic we are using.

Though we have no copy of $\mathbb{R}$ at hand, we can partially rectify this by using standard coding techniques, to get a more substantial domain of mathematical objects to work with. Indeed in the context of our sequence $\mathbb{N}$, we can code talk of a ring $\mathbb{Z}$ and then a field $\mathbb{Q}$: this is done by defining a pairing function, and then taking certain pairs of natural numbers to represent integers, and then certain pairs of integers to represent rationals.[1] In this coding, we need to be careful to distinguish elements of $\mathbb{Q}$ from elements of $\mathbb{N}$, with for instance $\frac{1}{1}$ as a rational perhaps not being coded as the natural number 1. We will write uncoded elements of $\mathbb{N}$ with subscripts $0_{\mathbb{N}}$, $1_{\mathbb{N}}$, and write whole rationals as fractions $\frac{0}{1}$, $\frac{1}{1}$ to avoid ambiguity. We will use this field $\mathbb{Q}$ to give the skeleton for the field structure on X, first pairing off elements of $\mathbb{Q}$ with elements of the countable

[1] For instance one might use $(n, 0)$ to represent $n \in \mathbb{Z}$, and $(n, 1)$ to represent $-n \in \mathbb{Z}$. Then one might use (a, b) to represent $\frac{a}{b} \in \mathbb{Q}$ where $a, b \in \mathbb{Z}$ and $b \neq 0$. For simplicity we will assume we are using fractions in lowest terms, i.e. whenever we write $\frac{c}{d}$ we signify the pair (a, b) where $a, b \in \mathbb{Z}$ are coprime, $b > 0$, and $ad = bc$.

dense sequence in X, as in the usual argument sketched above.

To do this, we need to be able to define a certain function from $\mathbb{Q}$ to X by recursion, a recursion in terms of all previously defined values of the function. This would normally be straightforward – if one was working in set theory, or even just with full second order logic – but because of our restricted logical toolbox, some care is needed (the double ancestral is not useful as we need a recursion in terms of all previously defined values of the function, rather than a simple primitive recursion). We will take a bit of time now to sketch how this can work in our setting, before returning to the main course of the argument.

The way such recursive functions are usually defined in the arithmetic context is by first coding up the ability to talk about finite sequences of objects. For this we want two functions $\mathrm{Len}(x)$ and $\beta(x, i)$, with the former giving the length of a sequence x and the latter being the function giving the i^{th} place of a sequence x. The key property we need is the ability to extend any sequence by adding a single element: so that for all k and n, if x has length k then there is y with length $k + 1$, where $\beta(y, i) = \beta(x, i)$ for $i \leqslant k$, and $\beta(y, k + 1) = n$. A standard way of defining functions with these properties (which works in our context) is seen in Buss (1998, pp. 92–94). From this it follows that if $\mathrm{Len}(x) = k$ and $l \leqslant k$ then there is y with $\mathrm{Len}(y) = l$ and $\beta(y, i) = \beta(x, i)$ for $i \leqslant l$. We write such a y as $x|_l$. With this in hand, we can obtain a scheme for recursive finite sequence definition. Suppose that $\phi(x, l, n)$ is any open formula such that for all x and l, if x is a finite sequence then there is (at least one) n such that $\phi(x, l, n)$ holds. Then we can prove that:

- For all k there is x such that for all l, if $0 \leqslant l < k$ then $\phi(x|_l, l, \beta(x, l + 1))$.

In other words each subsequent l^{th} element of the sequence x is defined in terms of ϕ from the index l and the previous elements of x. This is proved by induction on k, and importantly in our logical setting the inductive clause for R_S is phrased in plural logic, so inductions along $\mathbb{N}$ can be carried out for any formula we can define, including

formulae containing $R_<$. Thus this scheme for recursive finite sequence definition holds for all such open formulae $\phi(x, n)$, even if they involve $R_<$, or in other words, the relation $<_X$. If we put a uniqueness constraint on ϕ, namely that for all x and $l<$ if x is a finite sequence then there is a unique n such that $\phi(x, l, n)$ holds, then we obtain that any two finite sequences satisfying the above recursive finite sequence definition for this ϕ must have all the same values.

Then by using a pairing function, we obtain the ability to define a finite sequence $((a_1, b_1), \ldots (a_n, b_n))$ of pairs of natural numbers, with the value of each (a_i, b_i) defined in terms of the previous values $((a_1, b_1), \ldots (a_{i-1}, b_{i-1}))$. In particular we can use pairs of the form (i, b_i) to represent the values of a function f, giving us the ability to define a function on $\mathbb{N}$ by recursion on its values (for a well defined function on all of $\mathbb{N}$, we need the just mentioned uniqueness constraint on ϕ). Since the above scheme for recursive finite sequence definition applies even to formulae involving $<_X$, we can define functions by recursion on their values even when this recursion involves the relation $<_X$.

One simple use of this is to obtain an enumeration of the rationals: by sending n to the $<_{\mathbb{N}}$-smallest rational not enumerated so far, we can define a function by recursion which gives a bijection between $\mathbb{N}$ and $\mathbb{Q}$. We will write the value of this function at n by q_n. Combining this enumeration with the above ability to define functions along $\mathbb{N}$ by recursion gives us the ability to define functions along $\mathbb{Q}$ by recursion, defining where q_{k+1} is sent in terms of where $q_0 \ldots q_k$ are sent. Once again we are able to define such functions by a recursion involving the relation $<_X$ and the function α (the latter just being the identity function on $\mathbb{N}$).

With this in hand we are ready to define our function from $\mathbb{Q}$ to X by recursion. First, suppose that $x \in X$. As discussed initially, there is $n \in \mathbb{N}$ such that $\alpha(n) \in (x, \infty)$. Thus we can let

$$\text{Above}(x) = \alpha(\min_{<_{\mathbb{N}}}\{n \in \mathbb{N} \mid \alpha(n) \in (x, \infty)\}).$$

Similarly we can let

$$\text{Below}(x) = \alpha(\min_{<_{\mathbb{N}}}\{n \in \mathbb{N} \mid \alpha(n) \in (\infty, x)\}),$$

and if $x <_X y$, we can let

$$\text{Between}(x, y) = \alpha(\min_{<_{\mathbb{N}}}\{n \in \mathbb{N} \mid \alpha(n) \in (x, y)\}).$$

Now we define our function f from $\mathbb{Q}$ to X. First, we define f to send the initial element q_0 of $\mathbb{Q}$ in the enumeration to $\alpha(0_{\mathbb{N}})$. Then if we have defined the function f for the elements $q_0, \ldots q_k$ of $\mathbb{Q}$, we define it for q_{k+1} by cases.

- If q_{k+1} is $<_{\mathbb{Q}}$-greater than all the $q_1, \ldots q_k$, then we define $f(q_{k+1})$ to take the value $\text{Above}(\max_{<_X}\{f(q_1), \ldots f(q_k)\})$
- If q_{k+1} is $<_{\mathbb{Q}}$-less than all the $q_1, \ldots q_k$, then we define $f(q_{k+1})$ to take the value $\text{Below}(\min_{<_X}\{f(q_1), \ldots f(q_k)\})$
- If we have $q_i <_{\mathbb{Q}} q_{k+1} <_{\mathbb{Q}} q_j$, where $l \leqslant k$ implies $q_l \leqslant_{\mathbb{Q}} q_i$ or $q_l \geqslant_{\mathbb{Q}} q_j$, then we define $f(q_{k+1})$ to be $\text{Between}(f(q_i), f(q_j))$

This recursive definition is legitimate, by the previous remarks (recalling that the function α is just the identity function, with the symbol introduced to make clear the different roles of elements of our sequence – either as natural numbers, or as elements of the dense subset of X).

It is immediate by $\mathbb{N}$-induction that for each $l \in \mathbb{N}$, if $k < l$ then if $q_k <_{\mathbb{Q}} q_l$ then $f(q_k) <_X f(q_l)$, and if $q_k >_{\mathbb{Q}} q_l$ then $f(q_k) > f(q_l)$. Thus the map f is injective and order preserving as a map $(\mathbb{Q}, <_{\mathbb{Q}}) \to (X, <_X)$.

Next we can argue that its image consists exactly of all the $\alpha(n)$. Obviously it only takes values amongst the $\alpha(n)$. Conversely, we can argue by induction on n that for

each n, every $\alpha(0_{\mathbb{N}}) \dots \alpha(n)$ is in the image of f. This is obviously true for $0_{\mathbb{N}}$. If true for n, then we can find k such that $\alpha(0_{\mathbb{N}}) \dots \alpha(n)$ are amongst $f(q_0) \dots f(q_k)$. Then if $\alpha(n+1)$ is also one of $f(q_0) \dots f(q_k)$ then we are done. Otherwise one of three cases obtain for $\alpha(n+1)$:

- $\alpha(n+1) >_X f(q_i)$ for $i \leqslant_{\mathbb{N}} k$
- $\alpha(n+1) <_X f(q_i)$ for $i \leqslant_{\mathbb{N}} k$
- There are $i, j \leqslant_{\mathbb{N}} k$ with $f(q_i) <_X \alpha(n+1) <_X f(q_j)$ and for all $l \leqslant_{\mathbb{N}} k$ we have $f(q_l) \leqslant_X f(q_i)$ or $f(q_l) \geqslant_X f(q_j)$

In the first case, if we let $p \geqslant k+1$ be minimal such that $q_p >_{\mathbb{Q}} q_i$ for all $i \leqslant_{\mathbb{M}} k$, then we have $\alpha(n+1) = \text{Above}(\max_{<_X}\{f(q_0) \dots f(q_{p-1})\})$ and so $\alpha(n+1) = f(q_p)$. The second case is similar. In the third case, if we let $p \geqslant k+1$ be minimal such that $q_i < q_p < q_j$ then we have $\alpha(n+1) = \text{Between}(f(q_i), f(q_j))$ and so $\alpha(n+1) = f(q_p)$. Thus we have proved the induction hypothesis for $n+1$, so are done by induction.

From here on we will be taking various supremums and infimums of pluralities of elements of X, using the completeness property. We will write such a supremum for instance as $\sup\{x \mid \phi\}$, though this is a supremum of a plurality, not a set. For instance since the values of $f(q)$ for $q \in \mathbb{Q}$ are dense in X, we have for all $x \in X$ that $x = \sup\{f(q) \mid f(q) <_X x\}$.

Now obtaining the field structure on X is fairly straightforward. The argument is very similar to that giving a field structure on Dedekind cuts of rationals. Here we use $q, r, \dots$ for variables ranging over $\mathbb{Q}$, and $x, y \dots$ for variables ranging over X, with it being understood that by eg $q \leqslant r$ we mean $q \leqslant_{\mathbb{Q}} r$, and by $x \leqslant y$ we mean $x \leqslant_X y$. We let 0_X be $f(\frac{0}{1})$, we define $x +_X y$ to be $\sup\{f(q +_{\mathbb{Q}} r) \mid f(q) < x,\ f(r) < y\}$, and we define $-_X x = \inf\{f(-_{\mathbb{Q}} q) \mid f(q) < x\}$. We will often just denote these by 0, $+$ and $-$ respectively, allowing the context to make clear which object or operation is intended.

It is easy to see that $+_X$ is commutative, and that if $x \geqslant x'$ and $y \geqslant y'$ then $x + y \geqslant y + y'$. It follows from $x = \sup\{f(q) \mid f(q) < x\}$ that 0_X is an identity element and that $f(q + r) = f(q) + f(r)$. For associativity we argue that if $q \in \mathbb{Q}$ then $f(q) < x + (y + z)$ iff there are r, u, v with $q <_{\mathbb{Q}} r + \mathbb{Q}u +_{\mathbb{Q}} v$ and $f(r) <_X x$, $f(u) <_X y$, $f(v) <_X z$, which also holds iff $f(q) <_X (x + y) + z$, so that $f(q) <_X (x + y) + z$ iff $f(q) <_X x + (y + z)$, and thus $x + (y + z) = (x + y) + z$.

Now we have $f(q) \leqslant -x$ iff $f(q)$ is a lower bound for $\{f(-r) \mid f(r) < x\}$, iff whenever $f(r) < x$ we have $f(q) \leqslant f(-r)$, iff whenever $f(r) < x$ we have $q \leqslant -r$, iff whenever $f(r) < x$ we have $-q \geqslant r$, iff whenever $f(r) < x$ we have $f(-q) \geqslant f(r)$, iff $f(-q) \geqslant x$. Thus if $f(q) < x$ and $f(r) < -x$ then $f(q) < x \leqslant f(-r)$, and so $q < -r$, so $q + r < 0$, so $f(q + r) < 0$. Thus $0 \geqslant \sup\{f(q + r) \mid f(q) < x,\ f(r) < -x\} = x + (-x)$. Conversely if $f(s) < 0$ then $s > 0_{\mathbb{Q}}$ and we can find q with $f(q) < x < f(q - \frac{s}{2})$, and so $f(-q + \frac{s}{2}) \leqslant -x$ by the first fact noted. Thus $f(-q + \frac{3s}{4}) < -x$ and so $\sup\{f(q + r) \mid f(q) < x,\ f(r) < -x\} \geqslant f(q + (-q + \frac{3s}{4})) = f(\frac{3s}{4}) > f(s)$. Thus $\sup\{f(q+r) \mid f(q) < x,\ f(r) < -x\} \geqslant 0_X$, and so $0 = x + (-x)$.

Thus these definitions of 0_X, $+_X$ and $-_X$ give us a totally ordered group. We thus obtain standard group properties such as that $-(-x) = x$.

Next we define our ring structure. We let 1_X be $f(\frac{1}{1})$, and for $x, y > 0$ we define $x \times_X y$ to be the $\sup\{f(q \times_{\mathbb{Q}} r) \mid f(q) < x,\ f(r) < y\}$. We extend $x \times_X y$ for other arguments by cases:

- If $x = 0$ or $y = 0$ then we set $x \times_X y = 0$
- If $x \leqslant 0$ and $y \geqslant 0$ then we set $x \times_X y$ to be $-((-x) \times_X y)$
- If $x \geqslant 0$ and $y \leqslant 0$ then we set $x \times_X y$ to be $-(x \times_X (-y))$
- If $x \leqslant 0$ and $y \leqslant 0$ then we set $x \times_X y$ to be $(-x) \times_X (-y)$

It is easy to see that this gives a well defined function. It is also easy to see that this function is commutative, and that it has identity element 1_X. It is also easy to check

that for all y, $(-x) \times_X y = -(x \times_X y)$. In particular, $(-1) \times_X y = -y$. It is also obvious that if $x, y > 0$ then $x \times y > 0$.

Next we want to argue for associativity – that for x, y, z we have $x \times (y \times x) = (x \times y) \times z$. That this holds for strictly positive x, y, z can be argued in a similar manner to the case of addition. That it holds whenever one of x, y, z is 0 or 1 is easy, and that it holds whenever x, y or z is -1 follows from the above noted properties of negation. Then we can define the operation $x \mapsto x^n$ for $n \in \mathbb{N}$ as usual, where $x^0 = 1$ and $x^{n+1} = x \times_X x^n$.[2] We have $(-1)^2 = 1$, so by induction if $x \in \{1, -1\}$ then the only values that x^n takes are in $\{1, -1\}$, for which the commutative and associative properties of multiplication hold, so that $x^{n+m} = x^n \times x^m$ and $(x^n)^m = x^{nm}$ as usual. Now given $x \in X \setminus \{0\}$ there is a unique $a_x \in \{0, 1\}$ and a unique $x' > 0$ such that $x = (-1)^{a_x} x'$. These satisfy $(-1)^{a_{x \times y}} = (-1)^{a_x}(-1)^{a_y} = (-1)^{a_x + a_y}$, and $(x \times y)' = x' \times y'$. Then given $x, y, z \neq 0$ we have

$$\begin{aligned} & x \times (y \times z) \\ &= (-1^{a_x} \times x') \times ((-1)^{a_{y \times z}} \times (y \times z)') \\ &= ((-1)^{a_x} \times (-1)^{a_y} \times (-1)^{a_z}) \times (x' \times (y' \times z')) \\ &= ((-1)^{a_x} \times (-1)^{a_y} \times (-1)^{a_z}) \times ((x' \times y') \times z') \end{aligned}$$

(since x', y', z' are positive)

$$\begin{aligned} &= ((-1)^{a_{x \times y}} \times (x \times y)') \times ((-1)^{a_z} \times z') \\ &= (x \times y) \times z. \end{aligned}$$

Thus multiplication is associative.

One can prove that $(x + y) \times z = x \times z + y \times z$, for the case where $(x + y)$ and z are

[2] This can be defined using the double ancestral, though we will only need it for $x = 1, (-1)$ which are technically elements of $\mathbb{N}$ and so recursion on natural numbers suffices.

positive, in similar fashion to the proof that addition is associative. Then if any of x, y, z are 0 this identity is immediate. For $z = 1$ the identity is immediate, and for $z = (-1)$ it follows from properties of negation, so the identity holds when z is any power of (-1). Then for $(x + y), z \neq 0$ we have

$$(x + y) \times z = (-1)^{a_z + a_{x+y}} \times ((-1)^{a_{x+y}}(x + y) \times z')$$
$$= (-1)^{a_z + a_{x+y}} \times (((-1)^{a_{x+y}} \times x + (-1)^{a_{x+y}} \times y) \times z')$$

(from the identity for powers of (-1))

$$= (-1)^{a_z + a_{x+y}} \times ((-1)^{a_{x+y}} \times x \times z' + (-1)^{a_{x+y}} \times y \times z')$$

(by the identity for positive $(x + y), z$)

$$= (-1)^{a_z + a_{x+y}} \times (-1)^{a_{x+y}} \times x \times z' + (-1)^{a_z + a_{x+y}} \times (-1)^{a_{x+y}} \times y \times z'$$
$$= (-1)^{a_z} \times x \times z' + (-1)^{a_z} \times y \times z'$$
$$= x \times z + y \times z.$$

Thus we have an ordered ring structure, and all we need are multiplicative inverses. For $x > 0$ we define x^{-1} to be $\sup\{f(\frac{1}{q}) \mid f(q) > x\}$. First, if $f(r) < 1$ then pick q with $x < f(q) < x \times f(\frac{1}{r})$, and then we have by definition $f(\frac{1}{q}) \leqslant x^{-1}$ and so

$$x \times x^{-1} \geqslant x \times f(1/q) = x \times f(1/q) \times f(r \times (1/r))$$
$$= x \times f(1/r) \times f(r) \times f(1/q) > f(q) \times f(r) \times f(1/q) = f(r).$$

Since this holds for all $f(r) < 1$ we have $x \times x^{-1} \geqslant 1$. For the converse, given $f(r) > 1$ we can pick $u > 0$ with $f(\frac{1}{r}) \times x < f(u) < x$. Since $x > f(u)$, if q satisfies $f(q) > x$ then

$f(\frac{1}{u}) > f(\frac{1}{q})$, and so $f(\frac{1}{u}) \geqslant x^{-1}$. Thus

$$\begin{aligned} x \cdot x^{-1} &\leqslant x \times f(1/u) = x \times f(1/r) \times f(r) \times f(1/u) \\ &< f(u) \times f(r) \times f(1/u) = f(r). \end{aligned}$$

Since this holds for all $f(r) > 1$ we have $x \times x^{-1} \leqslant 1$, and thus $x \times x^{-1} = 1$. Thus this ordered ring structure has multiplicative inverses, so is an ordered field; which is complete, by the completeness property of $<_X$.

Appendix D

Interpreting mathematics in terms of a complete ordered field

Here we show how one can interpret more mathematics – including much of what is normally thought of as making up real analysis – in terms of a complete ordered field, as defined in section VI.4. That definition uses plural logic for the completeness axiom, and we also have double ancestral logic available, allowing definitions by primitive recursion. We will see how one can interpret talk of relation variables over a complete ordered field, by defining a pairing function; and how to interpret talk of sequences of elements of the field, and thus of sets like $\mathbb{R}^k$ for $k \in \mathbb{R}$. This work will allow us to show how various analytic properties are definable in this context, such as the property of a plurality being analytic, or $\mathbf{\Sigma}^1_2$, or having the Baire property, or the perfect set property, or being Lebesgue measurable; and properties concerning the determinary of infinite games. We will write $\mathbb{R}$ for the plurality of objects making up our complete ordered field, with $\mathbb{N}$ its plurality of natural numbers. We will use set builder notion $\{x \mid \phi\}$ to denote the plurality of objects satisfying ϕ.

We start by sketching how to define a pairing function for our complete ordered field, allowing us to simulate the use of relation variables. For each real x and natural

number n we can let $f_2(x, n)$ be the largest rational of the form $\frac{a}{2^n}$ which is at most x (using primitive recursion to define the function $n \mapsto 2^n$). Using these we can define the n^{th} place $g_2(x, n)$ in the binary expansion of x (chosen to be the terminating binary expansion if x is a dyadic rational). Then given x and y, by using a pairing function on their integer parts and interleaving their binary expansions appropriately we can define their pair $\rho(x, y)$, uniquely determined by x and y. Thus we will take it for granted that we have a pairing function, and can simulate the use of relation variables.

We want a good way of phrasing talk of sequences. We can define what it is for a binary relation on $\mathbb{R}$ to be a function $\mathbb{N} \to \mathbb{R}$; however we want to identify these relations on $\mathbb{R}$ with certain reals, via an injection $\mathbb{R}^{\mathbb{N}} \to \mathbb{R}$, so that we can talk about pluralities of them. There are various ways to define such an injection, normally using either base-n expansions for some n or continued fractions, and these can be straightforwardly expressed in the theory (the case of base-n expansions was sketched above). So we will take it for granted that we can pick an open formula $\Phi(F, x)$ where F is a binary relation variable, such that for all F if F defines a function $\mathbb{N} \to \mathbb{R}$ then there is a unique x such that $\Phi(F, x)$, and such that for all x there is at most one such F (and such that there only exists an x such that $\Phi(F, x)$ if F does define a function $\mathbb{N} \to \mathbb{R}$).

We will from now on use $\mathbb{R}^{\mathbb{N}}$ to denote the "image" of Φ. If $y \in \mathbb{R}^{\mathbb{N}}$ then we for $n \in \mathbb{N}$ we let $\eta(n, y)$ be the n^{th} place of y, i.e. $F(n)$ where F satisfies $\Phi(F, y)$.

Using these ingredients we can define what it is for a plurality of reals to be analytic. We start by for natural numbers $k \in \mathbb{N}$ defining $\mathbb{R}^k$ to be

$$\{\rho(p, k) \mid p \in \mathbb{R}^{\mathbb{N}} \wedge \forall i \geqslant k\, (\eta(i, p) = 0)\}$$

and for $z \in \mathbb{R}^k$ and $i = 1 \ldots k$ we let z_i be $\eta(i, z')$ where $z = \rho(z', z'')$. This gives us disjoint pluralities which can play the role of finite powers of $\mathbb{R}$. We can define the usual distance function on each $\mathbb{R}^k$, and thus define what it is for a subplurality of $\mathbb{R}^k$ to be

an "open plurality" (analogous to the usual notion of open set). We can also define the plurality $\mathbb{N}^{\mathbb{N}}$ contained in $\mathbb{R}^{\mathbb{N}}$, and define which subpluralities of $\mathbb{N}^{\mathbb{N}}$ are open (under the product topology – the usual topology on ω^{ω}). We can then define for each k what it is for a relation R to actually be a continuous function from $\mathbb{N}^{\mathbb{N}}$ to $\mathbb{R}^k$. Thus we can define the analytic subpluralities of $\mathbb{R}^k$: they are those which are either empty, or are the image of a continuous function from $\mathbb{N}^{\mathbb{N}}$ to $\mathbb{R}^k$.

We use the symbol $\mathbf{\Sigma}^1_1$ to apply to the analytic pluralities (analytic subpluralities of $\mathbb{R}^k$ for some k). We can define what it is for a plurality to be $\mathbf{\Pi}^1_1$ in the usual way: a subplurality X of $\mathbb{R}^k$ is $\mathbf{\Pi}^1_1$ if $\mathbb{R}^k \setminus X$ is $\mathbf{\Sigma}^1_1$. Then we can go on to define the $\mathbf{\Sigma}^1_2$ pluralities. A subplurality X of $\mathbb{R}^k$ is $\mathbf{\Sigma}^1_2$ if there is a $\mathbf{\Pi}^1_1$ subplurality Y of $\mathbb{R}^{k+1}$ such that X is $\{p \mid \exists q \in Y\, (i \leqslant k \to p_i = q_i)\}$.

We will now abuse notation slightly. We are really interested in the $\mathbf{\Sigma}^1_2$ subpluralities of $\mathbb{R}$, not the $\mathbf{\Sigma}^1_2$ subpluralities of $\mathbb{R}^1$; so we redefine $\mathbf{\Sigma}^1_2$ to apply to the images of the $\mathbf{\Sigma}^1_2$ subpluralities of $\mathbb{R}^1$ (by the old definition) under the obvious bijection between $\mathbb{R}$ and $\mathbb{R}^1$.

Next we will show how to define what it is for a plurality of reals to have the Baire property, the perfect set property, or be Lebesgue measurable.

To do this, we want to be able to talk about sequences of pluralities. We can just take these to be binary relations R such that for all x if there is a y such that $R(x, y)$ then x is in $\mathbb{N}$. If R is such a relation and $n \in \mathbb{N}$ then the n^{th} term of R is $\{y \mid R(n, y)\}$ and is denoted $\eta(n, R)$. We can also easily express what it is for a plurality to be countable: a plurality X is countable if there is a function F with domain $\mathbb{N}$, whose range is X.

With the ability to define sequences of pluralities, we can express what it is for a plurality to be meagre – a countable union of nowhere dense pluralities – and thus what it is for a plurality of reals to have the property of Baire, i.e. to have meagre symmetric difference with some open plurality. Next we can easily define what it is for a plurality to be perfect (i.e. closed and having no isolated points), and hence what it is for a plurality

to have the perfect set property, i.e. to be countable or contain a nonempty perfect subplurality. Finally using our ability to define sequences of pluralities we can define Lebesgue outer measure for pluralities. We let S apply to sequences of pluralities all of which are open intervals, and let $l(I)$ denote the length of an open interval I. Lebesgue outer measure λ^* is then given by

$$\lambda^*(X) = \inf\{ \sum_{k \geqslant 1} l(I_k) \mid S((I_k)_{k \in \mathbb{N}}) \text{ and } X \subseteq \bigcup_{k \geqslant 1} I_k \}.$$

With this we can define what it is for a plurality X to lie in the Lebesgue σ-algebra: it does so if for every plurality Y of reals,

$$\lambda^*(X) \geqslant \lambda^*(X \cap Y) + \lambda^*(X \cap Y^c).$$

To conclude, we consider questions of determinacy. We can define the plurality $\mathbb{N}^{\mathbb{N}}$ of reals, as seen previously. Subpluralities of $\mathbb{N}^{\mathbb{N}}$ can be thought of as target sets for games with a countably infinite number of moves: given a subplurality X of $\mathbb{N}^{\mathbb{N}}$, the first player in our game picks an element of $\mathbb{N}$, followed by the second player, then the first player, and so on, and if the sequence generated lies in X then the first player wins, but if it doesn't then the second player wins. We have already seen that we can define the pluralities $\mathbb{R}^k$, functions, and countable unions of pluralities in the analytic context, and accordingly we can take a strategy for the first player to be a function from $\bigcup_{k \geqslant 0} \mathbb{N}^{2k}$ to $\mathbb{N}$: the function's value on $(n_0, n_1 \ldots n_{2k})$ is the move the first player makes when the previous moves are $(n_0, n_1 \ldots n_{2k})$ (there is redundancy here since if n_1 is not the move the strategy dictates after n_0, then $(n_0, n_1 \ldots n_{2k})$ will never be played so the function's value on this sequence is irrelevant; this does not matter though). Similarly a strategy for the second player is a function from $\bigcup_{k \geqslant 0} \mathbb{N}^{2k+1}$ to $\mathbb{N}$. We can define what it is for a strategy to be a winning strategy in the usual way. Then the question of determinacy is the question for a subplurality X of $\mathbb{N}^{\mathbb{N}}$ of whether one of the players has a winning

strategy (if so, this subplurality defines a determined game).

Bibliography

Aberdein, Andrew and Ian J. Dove, eds. (2013). *The Argument of Mathematics*. Logic, Epistemology, and the Unity of Science. Springer Netherlands.

Alexander, J. W. (1923). "A Lemma on Systems of Knotted Curves". *Proceedings of the National Academy of Sciences of the United States of America* 9.3, pp. 93–95.

Alexander, J. W. and G. B. Briggs (1926). "On Types of Knotted Curves". *Annals of Mathematics* 28.1, pp. 562–586.

Aluffi, Paolo (July 30, 2009). *Algebra: Chapter 0*. Providence, R.I: American Mathematical Society. 713 pp.

Andersen, Line Edslev (Apr. 3, 2017). "On the Nature and Role of Peer Review in Mathematics". *Accountability in Research* 24.3, pp. 177–192.

Antonutti Marfori, Marianna (2010). "Informal Proofs and Mathematical Rigour". *Studia Logica: An International Journal for Symbolic Logic* 96.2, pp. 261–272.

Artin, Michael (Apr. 24, 1991). *Algebra*. Englewood Cliffs, N.J: Pearson. 618 pp.

Aschbacher, Michael (Oct. 15, 2005). "Highly complex proofs and implications of such proofs". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 363.1835, pp. 2401–2406.

Atiyah, Michael et al. (1994). "Responses to: A. Jaffe and F. Quinn, Theoretical mathematics: toward a cultural synthesis of mathematics and theoretical physics [Bull. Amer. Math. Soc. (N.S.) 29 (1993), no. 1, 113; MR1202292 (94h:00007)]". *Bulletin of the American Mathematical Society* 30.2, pp. 178–207.

Avigad, Jeremy (Dec. 1, 2011). "Understanding Proofs". In: *The Philosophy of Mathematical Practice*. Ed. by Paolo Mancosu. Oxford, New York: Oxford University Press, pp. 317–353.

Avron, Arnon (2003). "Transitive Closure and the Mechanization of Mathematics". In: *Thirty Five Years of Automating Mathematics*. Applied Logic Series. Springer, Dordrecht, pp. 149–171.

Awodey, Steve (Sept. 1, 1996). "Structure in Mathematics and Logic: A Categorical Perspective". *Philosophia Mathematica* 4.3, pp. 209–237.

— (Feb. 1, 2004). "An Answer to Hellman's Question: Does Category Theory Provide a Framework for Mathematical Structuralism?". *Philosophia Mathematica* 12.1, pp. 54–64.

— (June 17, 2010). *Category Theory*. Oxford ; New York: OUP Oxford. 328 pp.

Axler, Sheldon (Aug. 1, 1997). *Linear Algebra Done Right*. New York: Springer. 251 pp.

Azzouni, Jody (June 1, 2004). "The Derivation-Indicator View of Mathematical Practice". *Philosophia Mathematica* 12.2, pp. 81–106.

— (June 1, 2013). "The Relationship of Derivations in Artificial Languages to Ordinary Rigorous Mathematical Proof". *Philosophia Mathematica* 21.2, pp. 247–254.

Bak, Joseph and Donald J. Newman (2010). *Complex Analysis*. Undergraduate Texts in Mathematics. New York: Springer-Verlag.

Balaguer, Mark (July 19, 2001). *Platonism and Anti-Platonism in Mathematics*. Oxford: Oxford University Press. 240 pp.

Banach, S. (Mar. 1, 1987). *Theory of Linear Operations*. Elsevier. 249 pp.

Bartle, Robert Gardner and Donald R. Sherbert (2000). *Introduction to real analysis*. John Wiley & Sons Canada, Limited. 416 pp.

Berkeley, George (1999). "The analyst". In: *From Kant to Hilbert Volume 1: A Source Book in the Foundations of Mathematics*. Ed. by William Bragg Ewald. OUP Oxford, pp. 60–92.

Boolos, George (1971). "The Iterative Conception of Set". *The Journal of Philosophy* 68.8, pp. 215–231.

— (1984). "To Be is to be a Value of a Variable (or to be Some Values of Some Variables)". *The Journal of Philosophy* 81.8, pp. 430–449.

— (1985). "Nominalist Platonism". *The Philosophical Review* 94.3, pp. 327–344.

Brown, Bryson and Graham Priest (Aug. 1, 2004). "Chunk and Permeate, a Paraconsistent Inference Strategy. Part I: The Infinitesimal Calculus". *Journal of Philosophical Logic* 33.4, pp. 379–388.

Burde, Gerhard and Heiner Zieschang (Dec. 16, 2002). *Knots*. Berlin ; New York: de Gruyter. 572 pp.

Burgess, John P. (Oct. 1, 2004). "E Pluribus Unum: Plural Logic and Set Theory". *Philosophia Mathematica* 12.3, pp. 193–221.

— (July 5, 2005). *Fixing Frege*. Ed. by Harry G. Frankfurt. Princeton, N.J: Princeton University Press. 270 pp.

— (Feb. 12, 2015). *Rigor and Structure*. Oxford, United Kingdom: OUP Oxford. 240 pp.

Burgess, John P. and Gideon Rosen (Jan. 16, 1997). *A Subject With No Object: Strategies for Nominalistic Interpretation of Mathematics*. Oxford, New York: Oxford University Press. 272 pp.

Buss, S. R., ed. (July 9, 1998). *Handbook of Proof Theory*. New York: Elsevier Science. 810 pp.

Celluci, Carlo (Feb. 2009). "Why Proof? What is a Proof?" In: *Deduction, Computation, Experiment*. Ed. by Rossella Lupacchini and Giovanna Corsi. Springer Verlag Gmbh, pp. 1–27.

Chihara, Charles S. (June 6, 1991). *Constructibility and Mathematical Existence*. Oxford University Press.

Conway, John B. (1978). *Functions of One Complex Variable I*. Graduate Texts in Mathematics, Functions of One Complex Variable. New York: Springer-Verlag.

Corcoran, J. (1980). "On Definitional Equivalence and Related Topics". *History and Philosophy of Logic* 1.1, p. 231.

Cori, René and Daniel Lascar (Nov. 9, 2000). *Mathematical Logic Part 1: Propositional Calculus, Boolean Algebras & Predicate Calculus: A Course with Exercises*. Trans. by Donald H. Pelletier. Oxford ; New York: Oxford University Press, USA. 360 pp.

Corry, Leo (Dec. 6, 2012). *Modern Algebra and the Rise of Mathematical Structures*. Birkhäuser. 463 pp.

Coulston, Charles, ed. (1970). *Dictionary of Scientific Biography Volume II: Hans Berger - Christoph Buys Ballot*. Charles Scribner's Sons.

Dalvit, Ester (2012). *A journey through the mathematical theory of braids*.

De Toffoli, Silvia and Valeria Giardino (Aug. 1, 2014). "Forms and Roles of Diagrams in Knot Theory". *Erkenntnis* 79.4, pp. 829–842.

— (2015). "An Inquiry into the Practice of Proving in Low-Dimensional Topology". In: *From Logic to Practice: Italian Studies in the Philosophy of Mathematics*. Ed. by Gabriele Lolli, Marco Panza, and Giorgio Venturi. Boston Studies in the Philosophy and History of Science. Springer International Publishing, pp. 315–336.

— (May 26, 2016). "Envisioning Transformations - The Practice of Topology". In: *Mathematical Cultures: The London Meetings 2012-2014*. Ed. by Brendan Larvor. New York, NY: Birkhäuser, pp. 25–50.

Detlefsen, Michael (Jan. 22, 2009). "Proof: Its Nature and Significance". In: *Proof and Other Dilemmas: Mathematics and Philosophy*. Ed. by Bonnie Gold and Roger Simons. Washington, D.C.: Cambridge University Press, pp. 3–32.

Eisenbud, David (Mar. 30, 1995). *Commutative Algebra: With a View Toward Algebraic Geometry*. Springer Science & Business Media. 822 pp.

Eisenbud, David and Joe Harris (2000). *The Geometry of Schemes*. Graduate Texts in Mathematics. New York: Springer-Verlag.

Epp, Susanna (2003). "The Role of Logic in Teaching Proof". *The American Mathematical Monthly* 110.10, pp. 886–899.

— (2009). "Proof Issues with Existential Quantification". In: *Proof and Proving in Mathematics Education: The 19th ICMI Study*. Ed. by Fou-Lai Lin et al. Vol. 1, pp. 154–159.

Erdös, P. (Apr. 1947). "Some remarks on the theory of graphs". *Bulletin of the American Mathematical Society* 53.4, pp. 292–294.

Field, Hartry (Dec. 21, 1980). *Science Without Numbers*. Princeton, N.J: Princeton University Press. 144 pp.

— (Aug. 10, 1989). *Realism, Mathematics and Modality*. Oxford, UK ; New York, NY, USA: Wiley-Blackwell. 304 pp.

Fremlin, David (Nov. 8, 2010). *Measure Theory Volume 1*. Vol. 1. 5 vols. Torres Fremlin. 102 pp.

Fulton, William and Joe Harris (1991). *Representation theory: a first course*. Springer-Verlag. 576 pp.

Gentzen, Gerhard (1969). *The Collected Papers of Gerhard Gentzen*. Amsterdam: North-Holland Pub. Co.

Goethe, Norma B. and Michèle Friend (Nov. 1, 2010). "Confronting Ideals of Proof with the Ways of Proving of the Research Mathematician". *Studia Logica* 96.2, pp. 273–288.

Gowers, Timothy (Aug. 22, 2002). *Mathematics: A Very Short Introduction*. Oxford ; New York: OUP Oxford. 160 pp.

Gowers, Timothy, June Barrow-Green, and Imre Leader, eds. (Sept. 28, 2008). *The Princeton Companion to Mathematics*. Princeton: Princeton University Press. 1056 pp.

Grcar, Joseph (2013). "Errors and Corrections in Mathematics Literature". *Notices of the American Mathematical Society* 60.4, pp. 418–425.

Guillemin, Victor and Alan Pollack (Aug. 24, 1974). *Differential Topology*. Englewood Cliffs, N.J: Prentice Hall. 222 pp.

Hale, Bob (2005). "Real Numbers and Set Theory: Extending the Neo-Fregean Programme beyond Arithmetic". *Synthese* 147.1, pp. 21–41.

Hale, Bob and Crispin Wright (Jan. 15, 2004). *The Reason's Proper Study: Essays towards a Neo-Fregean Philosophy of Mathematics*. Oxford: Oxford University Press, USA. 470 pp.

Hales, Thomas C. (2007). "Jordan's Proof of the Jordan Curve Theorem". *Studies in Logic, Grammar and Rhetoric* 10.23.

Halmos, Paul R. (Aug. 17, 2011). *Naive Set Theory*. Mansfield (CT): Martino Fine Books. 114 pp.

Hartshorne, Robin (1977). *Algebraic Geometry*. Graduate Texts in Mathematics. New York: Springer-Verlag.

Hatcher; Allen (Dec. 3, 2001). *Algebraic Topology by Allen Hatcher*. Cambridge University Press.

Heck, Richard Kimberly (Sept. 29, 2011). *Frege's Theorem*. OUP Oxford. 322 pp.

Hellman, Geoffrey (1993). *Mathematics Without Numbers: Towards a Modal-structural Interpretation*. Oxford University Press. 172 pp.

Hersh, Reuben (Aug. 21, 1997). *What Is Mathematics, Really?* Oxford University Press. 369 pp.

Hilbert, David (July 1, 1990). "On the infinite". In: *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931*. Ed. by Jean van Heijenoort. Cambridge, Mass: Harvard University Press, pp. 367–392.

Hirsch, Morris W. (1976). *Differential Topology*. Graduate Texts in Mathematics. New York: Springer-Verlag.

Honsberger, R. (Aug. 30, 1978). *Mathematical Gems: Pt. 1*. Washington, DC: John Wiley & Sons. 187 pp.

Isaacson, Daniel (Jan. 1, 1987). "Arithmetical Truth and Hidden Higher Order Concepts". In: *Logic Colloquium '85*. Ed. by The Paris Logic Group. Elsevier, pp. 147–169.

— (Jan. 9, 1992). "Some Considerations on Arithmetical Truth and the -Rule". In: *Proof, Logic and Formalization*. Ed. by Michael Detlefsen. London ; New York: Routledge, pp. 94–138.

Jaffe, Arthur and Frank Quinn (1993). "Theoretical mathematics: toward a cultural synthesis of mathematics and theoretical physics". *Bulletin of the American Mathematical Society* 29.1, pp. 1–13.

James, Gordon and Martin Liebeck (Oct. 18, 2001). *Representations and Characters of Groups*. Cambridge, UK ; New York, NY: Cambridge University Press. 468 pp.

Jech, Thomas (Mar. 21, 2006). *Set Theory: The Third Millennium Edition, revised and expanded*. Berlin ; New York: Springer. 796 pp.

Jones, Vaughan (1998). "A credo of sorts". In: *Truth in Mathematics*. Ed. by Harold G. Dales and Gianluigi Oliveri. Clarendon Press, pp. 203–214.

Jordan, Camille (1887). *Cours d'analyse de l'ecole polytechnique*. 3 vols.

Kanamori, Akihiro (May 20, 1997). *The Higher Infinite: Large Cardinals in Set Theory from Their Beginnings*. Berlin ; New York: Springer. 536 pp.

Kitcher, Philip (Apr. 11, 1985). *The Nature of Mathematical Knowledge*. New York: Oxford University Press. 300 pp.

Kline, Morris (Mar. 1, 1990a). *Mathematical Thought from Ancient to Modern Times, Vol. 1*. Vol. 1. 3 vols. New York: Oxford University Press. 432 pp.

— (Aug. 16, 1990b). *Mathematical Thought from Ancient to Modern Times, Vol. 3*. Vol. 3. 3 vols. New York: Oxford University Press. 444 pp.

Konyndyk, Kenneth (May 31, 1986). *Introductory Modal Logic*. Notre Dame, Ind: University of Notre Dame Press. 144 pp.

Krantz, Steven G. (2011). *The Proof is in the Pudding: The Changing Nature of Mathematical Proof*. New York: Springer-Verlag.

Ladyman, James and Stuart Presnell (June 1, 2018). "Does Homotopy Type Theory Provide a Foundation for Mathematics?" *The British Journal for the Philosophy of Science* 69.2, pp. 377–420.

Larvor, Brendan (July 1, 2012). "How to think about informal proofs". *Synthese* 187.2, pp. 715–730.

— (July 2019). "From Euclidean geometry to knots and nets". *Synthese* 196.7, pp. 2715–2736.

Lavine, Shaughan (Jan. 13, 1998). *Understanding the Infinite*. Cambridge, Mass.: Harvard University Press. 384 pp.

Lee, John (Aug. 26, 2012). *Introduction to Smooth Manifolds*. New York ; London: Springer. 726 pp.

Lee, John M. (2000). *Introduction to Topological Manifolds*. Graduate Texts in Mathematics. New York: Springer-Verlag.

Leitgeb, Hannes (Sept. 29, 2009). "On Formal and Informal Provability". In: *New Waves in Philosophy of Mathematics*. Ed. by O. Bueno and Ø Linnebo. Basingstoke ; New York: Palgrave Macmillan, pp. 263–299.

Leng, Mary (Apr. 22, 2010). *Mathematics and Reality*. Oxford: OUP Oxford. 290 pp.

Linnebo, Øystein (2010). "Pluralities and Sets". *The Journal of Philosophy* 107.3, pp. 144–164.

— (June 2013). "The Potential Hierachy of Sets". *The Review of Symbolic Logic* 6.2, pp. 205–228.

— (June 14, 2018). *Thin Objects: An Abstractionist Account*. Oxford, New York: Oxford University Press. 256 pp.

Linnebo, Øystein and Richard Pettigrew (Oct. 1, 2011). "Category Theory as an Autonomous Foundation". *Philosophia Mathematica* 19.3, pp. 227–254.

MacLane, Saunders (Sept. 25, 1998). *Categories for the Working Mathematician*. New York: Springer. 332 pp.

MacLane, Saunders and Ieke Moerdijk (Oct. 27, 1994). *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Springer Science & Business Media. 651 pp.

Maddy, Penelope (1988a). "Believing the Axioms. I". *The Journal of Symbolic Logic* 53.2, pp. 481–511.

— (1988b). "Believing the Axioms. II". *The Journal of Symbolic Logic* 53.3, pp. 736–764.

— (1997). *Naturalism in Mathematics*. Clarendon Press. 265 pp.

— (Jan. 27, 2011). *Defending the Axioms: On the Philosophical Foundations of Set Theory*. Oxford, New York: Oxford University Press. 162 pp.

Mancosu, Paolo (Aug. 1, 1989). "The metaphysics of the calculus: A foundational debate in the Paris Academy of Sciences, 17001706". *Historia Mathematica* 16.3, pp. 224–248.

Manders, Kenneth (June 19, 2008). "The Euclidean Diagram". In: *The Philosophy of Mathematical Practice*. Ed. by Paolo Mancosu. OUP Oxford, pp. 80–133.

Martin, Donald A. (1998). "Mathematical Evidence". In: *Truth in Mathematics*. Ed. by Harold G. Dales and Gianluigi Oliveri. Clarendon Press, pp. 215–231.

Martin, R. M. (1943). "A Homogeneous System for Formal Logic". *The Journal of Symbolic Logic* 8.1, pp. 1–23.

— (1949). "A Note on Nominalism and Recursive Functions". *The Journal of Symbolic Logic* 14.1, pp. 27–31.

McLarty, Colin (2004). "Exploring Categorical Structuralism". *Philosophia Mathematica* 12.1, pp. 37–53.

— (Feb. 1, 2012). "Categorical Foundations and Mathematical Practice". *Philosophia Mathematica* 20.1, pp. 111–113.

Morgan, John and Gang Tian (Aug. 30, 2007). *Ricci Flow and the Poincare Conjecture*. Providence, RI: American Mathematical Society. 521 pp.

Muller, F. A. (2004). "The Implicit Definition of the Set-Concept". *Synthese* 138.3, pp. 417–451.

Mumma, John (July 1, 2010). "Proofs, pictures, and Euclid". *Synthese* 175.2, pp. 255–287.

Munkres, James (Jan. 7, 2000). *Topology*. Upper Saddle River, NJ: Pearson. 537 pp.

Nathanson, Melvyn (2008). "Desperately seeking mathematical truth". *Notices of the American Mathematical Society* 55.7, p. 773.

Nelson, Edward (Nov. 1977). "Internal set theory: A new approach to nonstandard analysis". *Bulletin of the American Mathematical Society* 83.6, pp. 1165–1198.

Niven, Ivan Morton, Avan Niven, and Herbert S. Zuckerman (Jan. 2, 1991). *An Introduction to the Theory of Numbers*. New York: John Wiley & Sons. 544 pp.

Oliver, Alex and Timothy Smiley (June 6, 2013). *Plural Logic*. Oxford: OUP Oxford. 352 pp.

Parsons, Charles (Dec. 24, 2007). *Mathematical Thought and its Objects*. Cambridge University Press. 400 pp.

Paseau, Alexander (Feb. 1, 2007). "Boolos on the justification of set theory". *Philosophia Mathematica* 15.1, pp. 30–53.

— (Dec. 1, 2015). "Knowledge of Mathematics without Proof". *The British Journal for the Philosophy of Science* 66.4, pp. 775–799.

Pelc, Andrzej (Feb. 1, 2009). "Why Do We Believe Theorems?" *Philosophia Mathematica* 17.1, pp. 84–94.

Perelman, Grisha (Nov. 11, 2002). "The entropy formula for the Ricci flow and its geometric applications". arXiv: `math/0211159`.

— (July 17, 2003a). "Finite extinction time for the solutions to the Ricci flow on certain three-manifolds". arXiv: `math/0307245`.

— (Mar. 10, 2003b). “Ricci flow with surgery on three-manifolds”. arXiv: `math/0303109`.

Potter, Michael (Mar. 11, 2004). *Set Theory and Its Philosophy: A Critical Introduction*. Oxford ; New York: Clarendon Press. 360 pp.

Prawitz, Dag (1965). *Natural Deduction: A Proof-Theoretical Study*. Dover Publications.

Rav, Yehuda (Feb. 1, 1999). “Why Do We Prove Theorems?” *Philosophia Mathematica* 7.1, pp. 5–41.

— (Oct. 1, 2007). “A Critique of a Formalist-Mechanist Version of the Justification of Arguments in Mathematicians’ Proof Practices”. *Philosophia Mathematica* 15.3, pp. 291–320.

Reidemeister, Kurt (Dec. 1, 1927). “Elementare Begründung der Knotentheorie”. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg* 5.1, pp. 24–32.

Robinson, Abraham (Jan. 8, 1996). *Non-standard Analysis*. Princeton, N.J: Princeton University Press. 320 pp.

Rowlett, Peter (July 2011). “The unplanned impact of mathematics”. *Nature* 475.7355, pp. 166–169.

Rudin, Walter (1987). *Real and complex analysis*. McGraw-Hill. 438 pp.

Schwartz, J. (1954). “The Formula for Change in Variables in a Multiple Integral”. *The American Mathematical Monthly* 61.2, pp. 81–85.

Shapiro, Stewart (1991). *Foundations Without Foundationalism: A Case for Second-Order Logic*. Oxford University Press.

— (Jan. 1, 1997). *Philosophy of Mathematics: Structure and Ontology*. Oxford: Oxford University Press USA. 296 pp.

Silverman, Joseph H. (Jan. 18, 2012). *A Friendly Introduction to Number Theory*. Boston: Pearson. 432 pp.

Smith, Peter (2008). “Ancestral Arithmetic and Isaacson’s Thesis”. *Analysis* 68.297, pp. 1–10.

Studd, J. P. (Oct. 1, 2013). "The Iterative Conception of Set". *Journal of Philosophical Logic* 42.5, pp. 697–725.

Sweeney, David John (Jan. 2, 2014). "Chunk and Permeate: The Infinitesimals of Isaac Newton". *History and Philosophy of Logic* 35.1, pp. 1–23.

Switzer, Robert M. (2002). *Algebraic Topology - Homotopy and Homology*. Classics in Mathematics. Berlin Heidelberg: Springer-Verlag.

Szemerédi, E. (1975). "On sets of integers containing no k elements in arithmetic progression". *Polska Akademia Nauk. Instytut Matematyczny. Acta Arithmetica* 27, pp. 199–245.

Tait, William W. (Jan. 1, 2005). *The Provenance of Pure Reason: Essays in the Philosophy of Mathematics and Its History*. Oxford ; New York: Oxford University Press. 348 pp.

Tanswell, Fenner (Oct. 1, 2015). "A Problem with the Dependence of Informal Proofs on Formal Proofs". *Philosophia Mathematica* 23.3, pp. 295–310.

Tao, Terence (2009). *Theres more to mathematics than rigour and proofs*. What's new.

Tatton-Brown, Oliver (2019a). "Primitive Recursion and Isaacson's Thesis". *Thought: A Journal of Philosophy* 8.1, pp. 4–15.

— (Dec. 13, 2019b). "Rigour and Intuition". *Erkenntnis*.

Taylor, John and Rowan Garnier (Apr. 7, 2014). *Understanding Mathematical Proof*. Boca Raton: Routledge. 416 pp.

Thurston, William P. (1994). "On proof and progress in mathematics". *Bulletin of the American Mathematical Society* 30.2, pp. 161–177.

— (Jan. 17, 1997). *Three-Dimensional Geometry and Topology, Vol. 1*. Ed. by Silvio Levy. Princeton, N.J: Princeton University Press. 328 pp.

Tieszen, Richard (Feb. 13, 1992). "What is a proof?" In: *Proof, Logic and Formalization*. Ed. by Michael Detlefsen. London ; New York: Routledge, pp. 57–76.

Tragesser, Robert (Feb. 13, 1992). “Three Insufficiently Attended to Aspects of Most Mathematical Proofs”. In: *Proof, Logic and Formalization*. Ed. by Michael Detlefsen. London ; New York: Routledge, pp. 162–198.

Univalent Foundations Program, The (2013). *Homotopy Type Theory: Univalent Foundations of Mathematics*.

Veblen, Oswald (1905). “Theory on Plane Curves in Non-Metrical Analysis Situs”. *Transactions of the American Mathematical Society* 6.1, pp. 83–98.

Velleman, Daniel J. (Apr. 27, 2006). *How to Prove It: A Structured Approach*. Cambridge ; New York: Cambridge University Press. 398 pp.

Walker, Russell C. (1974). *The Stone-ech Compactification*. Springer-Verlag. 350 pp.

Weir, Alan (Mar. 2016). “Informal Proof, Formal Proof, Formalism”. *The Review of Symbolic Logic* 9.1, pp. 23–43.

Wright, Crispin (Jan. 1, 1983). *Frege’s Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.